

**Some pages of this thesis may have been removed for copyright restrictions.**

If you have discovered material in AURA which is unlawful e.g. breaches copyright, (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please read our [Takedown Policy](#) and [contact the service](#) immediately

# A computational model of prosody for Yorùbá text-to-speech synthesis

ỌDÉTÚNJÍ ÀJÀDÍ ỌDÉJỌBÍ

Doctor of Philosophy

Supervisor: Dr. A. J. Beaumont

Co-Supervisor: Dr. S. H. S. Wong

ASTON UNIVERSITY

July 2005

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without proper acknowledgement.

# A computational model of prosody for Yorùbá text-to-speech synthesis

ỌDÉTÚNJÍ ÀJÀDÍ ỌDÉJỌBÍ

Doctor of Philosophy, 2005

## Abstract

This work examines prosody modelling for the Standard Yorùbá (SY) language in the context of computer text-to-speech synthesis applications. The thesis of this research is that it is possible to develop a practical prosody model by using appropriate computational tools and techniques which combines acoustic data with an encoding of the phonological and phonetic knowledge provided by experts. Our prosody model is conceptualised around a modular holistic framework. The framework is implemented using the Relational Tree (R-Tree) techniques (*Ehrich and Foith, 1976*). R-Tree is a sophisticated data structure that provides a multi-dimensional description of a waveform. A Skeletal Tree (S-Tree) is first generated using algorithms based on the tone phonological rules of SY. Subsequent steps update the S-Tree by computing the numerical values of the prosody dimensions.

To implement the intonation dimension, fuzzy control rules were developed based on data from native speakers of Yorùbá. The Classification And Regression Tree (CART) and the Fuzzy Decision Tree (FDT) techniques were tested in modelling the duration dimension. The FDT was selected based on its better performance.

An important feature of our R-Tree framework is its flexibility in that it facilitates the independent implementation of the different dimensions of prosody, i.e. duration and intonation, using different techniques and their subsequent integration.

Our approach provides us with a flexible and extendible model that can also be used to implement, study, and explain the theory behind aspects of the phenomena observed in speech prosody.

**Keywords:** Prosody modelling, speech synthesis, intonation modelling, duration modelling, fuzzy logic, fuzzy decision tree, relational tree

# Acknowledgements

I wish to thank the British people who, through the Commonwealth Scholarship Commission in the United Kingdom, awarded me the prestigious Commonwealth Scholarship. The generous award, efficiently administered by The British Council, allowed me to carry out the research reported in this thesis.

I thank the Ọbáfémi Awólówò University, Ilé-Ifè, Nigeria, for providing all the necessary supports throughout the duration of my study. I also thank my colleagues and students in the Computer Science and Engineering Department, Ọbáfémi Awólówò University, for bearing with my very long absence.

I thank Prof. J. O. A. Ayeni of the Computer Science Department, University of Lagos, Nigeria, for his contributions to the initial ideas that led to this research.

Many thanks to Prof. Robert Ladd of the University of Edinburgh, who facilitated my visit to the Centre for Speech Technology Research (CSTR), and Prof. Steve Renals, Dr. S. King and Dr. R. Clark who assisted and thought me during my research visit to the CSTR. Thanks to Dr. Mark Huckvale of the University College, London, for his insightful review of this work.

A lot of thanks to Dr. John Coleman of Oxford University who allowed me to use the Oxford University Phonetic Laboratory facilities and Prof. Greg Kochanski for his lectures and many assistance on the development of the Stem-ML model for Yorùbá.

I thank Prof. Dafydd Gibbon of the University of Bielefeld, Germany, for sharing his ideas and enlightening me at a very critical stage in this research. I also thank Prof. Akin Akinlabí of the Rutgers State University, USA, for helping me to get some of his publications.

My supervisor, Dr. A. J. Beaumont, has been a great source of education, guidance and encouragement. I sincerely appreciate the critical review, effective scrutiny, and constant feedback from my co-supervisor, Dr. S. H. S Wong. A lot of thanks due to S. M. Alamohoda for his personal and brotherly advise when the going was tough. A big thank you is due to all Aston University's computer science technical support staff, who helped to get things going as well as my colleagues in room MB306.

My wife, Foláké Ayòbámi, and my children, Olúrìnre and Olúmáyòmídé, as well as my extended family have been a great pillar of support and a source of encouragement.

# Contents

<b>I</b>	<b>Introduction</b>	<b>15</b>
<b>1</b>	<b>Prosody modelling: a synopsis</b>	<b>16</b>
1.1	Definition of terms . . . . .	17
1.1.1	An utterance . . . . .	17
1.1.2	Speech . . . . .	18
1.1.3	Text . . . . .	18
1.1.4	Fundamental frequency ( $f_0$ ), pitch, tone, and intonation . . . . .	18
1.1.5	Duration . . . . .	20
1.1.6	Prosody . . . . .	20
1.1.7	Linguistic . . . . .	21
1.2	Issues in prosody modelling . . . . .	22
1.3	Motivation . . . . .	23
1.3.1	Theoretical motivation . . . . .	24
1.3.2	Practical motivation . . . . .	26
1.4	Focus and scope of research . . . . .	28
1.4.1	Research context . . . . .	28
1.4.2	Research scope . . . . .	28
1.4.3	Empirical hypothesis . . . . .	30
1.5	Yorùbá: a brief introduction . . . . .	30
1.6	Thesis structure . . . . .	31
<b>II</b>	<b>Research Background and Literature Review</b>	<b>33</b>
<b>2</b>	<b>Background of speech synthesis technology</b>	<b>34</b>
2.1	A brief history of speech synthesis technology . . . . .	34
2.1.1	The manual and mechanical era . . . . .	34
2.1.2	Electrical and electronic era . . . . .	37
2.2	State-of-the-art in text-to-speech synthesis . . . . .	39
2.2.1	Modern TTS systems . . . . .	39
2.2.2	Review of High Level Synthesis (HLS) . . . . .	41
2.2.3	Review of techniques in Low Level Synthesis (LLS) . . . . .	45
2.3	Unit of speech synthesis . . . . .	52
2.4	Summary . . . . .	56

<b>3</b>	<b>The standard Yorùbá language</b>	<b>57</b>
3.1	Phonology . . . . .	57
3.2	Phonological structure of Yorùbá syllable . . . . .	58
	3.2.1 Syllable inventory . . . . .	61
3.3	SY tone phonology and phonetics . . . . .	62
	3.3.1 SY syllable and prosody . . . . .	63
3.4	The standard Yorùbá orthography . . . . .	64
	3.4.1 SY texts . . . . .	66
<b>4</b>	<b>A review of intonation models</b>	<b>70</b>
4.1	Data-driven approach to intonation and prosody modelling . . . . .	71
	4.1.1 Hidden Markov Models (HMM) based intonation models . . . . .	72
	4.1.2 Genetic Algorithms based intonation models . . . . .	75
	4.1.3 Neural Networks based intonation models . . . . .	75
4.2	Theoretical models of intonation and prosody . . . . .	78
	4.2.1 Pierrehumbert's model of intonation . . . . .	79
	4.2.2 Taylor's model . . . . .	82
	4.2.3 INTSINT . . . . .	85
	4.2.4 Fujisaki model . . . . .	87
	4.2.5 The Stem-ML model . . . . .	90
	4.2.6 Gibbon's computational model of intonation . . . . .	92
	4.2.7 Ladd's intonation model . . . . .	96
4.3	Summary . . . . .	99
<b>5</b>	<b>A review of duration models</b>	<b>101</b>
5.1	Rule-based duration model . . . . .	102
5.2	Mathematical equation-based duration models . . . . .	104
5.3	Data-driven duration models . . . . .	105
5.4	Other duration models . . . . .	108
5.5	Summary . . . . .	109
<b>III</b>	<b>Design Methodology</b>	<b>110</b>
<b>6</b>	<b>Design tools and techniques</b>	<b>111</b>
6.1	The relational tree . . . . .	111
6.2	Fuzzy logic and fuzzy control . . . . .	116
6.3	Classification And Regression Tree (CART) . . . . .	121
6.4	Fuzzy Decision Tree (FDT) . . . . .	125
6.5	Text markup in TTS . . . . .	127
6.6	Summary . . . . .	130
<b>7</b>	<b>Model conceptualisation and design</b>	<b>131</b>
7.1	Overview of the SY TTS system . . . . .	131
7.2	Motivation of our approach to modelling . . . . .	134
7.3	Overview of the synthesis strategy . . . . .	135
	7.3.1 The syllable overlay strategy . . . . .	139
7.4	Overview of our prosody modelling technique . . . . .	140

7.5	Abstract waveform generation . . . . .	142
7.6	Dimension computation . . . . .	143
7.6.1	Computing the intonation dimension . . . . .	143
7.6.2	Computing the duration dimension . . . . .	145
7.7	Integration of the dimension . . . . .	147
7.8	Summary . . . . .	147
<b>8</b>	<b>Research data</b>	<b>149</b>
8.1	Collection of text material . . . . .	149
8.2	Speech corpus . . . . .	152
8.2.1	Recording . . . . .	153
8.2.2	Recording equipments and environment . . . . .	153
8.2.3	Speech file annotation . . . . .	154
8.3	Summary . . . . .	157
<b>IV</b>	<b>Model Implementation</b>	<b>160</b>
<b>9</b>	<b>Standard Yorùbá prosody model design</b>	<b>161</b>
9.1	The domain of prosody in SY speech . . . . .	161
9.2	$f_0$ stylisation and standardisation . . . . .	165
9.3	Related works . . . . .	166
9.3.1	Linear function based model . . . . .	167
9.3.2	Quadratic model . . . . .	171
9.3.3	Cubic model . . . . .	172
9.3.4	Comparison of models . . . . .	173
9.4	SY $f_0$ stylisation . . . . .	175
9.4.1	Linear interpolation . . . . .	176
9.4.2	Quadratic interpolation . . . . .	178
9.4.3	Higher degree interpolations . . . . .	179
9.5	Evaluation of the stylisation functions . . . . .	181
9.5.1	Quantitative evaluation . . . . .	182
9.5.2	Qualitative evaluation . . . . .	183
9.5.3	Summary of stylisation experiment . . . . .	185
9.6	Standardisation . . . . .	186
9.7	Abstract intonation pattern generation . . . . .	188
9.7.1	Skeletal tree generation algorithm . . . . .	190
9.8	Skeletal tree generation for SY sentences . . . . .	192
9.9	Extension of the S-Tree algorithm . . . . .	195
9.10	Features of the S-Tree model . . . . .	201
9.11	Summary . . . . .	202
<b>10</b>	<b>Intonation modelling</b>	<b>204</b>
10.1	The training data . . . . .	205
10.2	Fuzzy model design . . . . .	206
10.2.1	Related models . . . . .	206
10.2.2	Model structure identification . . . . .	208
10.2.3	Generating intonation contour for an utterance . . . . .	219

10.3	Model evaluation . . . . .	223
10.3.1	Quantitative evaluation . . . . .	226
10.3.2	Qualitative evaluation . . . . .	228
10.3.3	Statistics difference calculation . . . . .	228
10.4	Summary . . . . .	234
<b>11</b>	<b>Duration modelling</b>	<b>235</b>
11.1	A preliminary analysis of factors affecting duration in SY . . . . .	237
11.1.1	Duration factor level and hierarchy . . . . .	238
11.2	Statistical analysis and observation . . . . .	239
11.2.1	Level 0 factors . . . . .	241
11.2.2	Levels 1–3 factors . . . . .	244
11.2.3	Pause duration . . . . .	247
11.3	SY syllable duration modelling . . . . .	247
11.3.1	Duration data normalisation . . . . .	248
11.4	FDT in duration modelling . . . . .	249
11.4.1	Problem formulation . . . . .	250
11.4.2	Fuzzy decision tree design . . . . .	251
11.4.3	Fuzzification of the input space . . . . .	252
11.4.4	Building the FDT . . . . .	255
11.4.5	Fuzzy decision tree pruning . . . . .	258
11.4.6	Applying FDT to duration modelling . . . . .	261
11.5	CART duration model . . . . .	262
11.6	Evaluation . . . . .	264
11.6.1	Qualitative evaluation . . . . .	266
11.6.2	Intelligibility evaluation . . . . .	267
11.6.3	Naturalness evaluation . . . . .	268
11.7	Summary . . . . .	269
<b>12</b>	<b>Prosody model implementation</b>	<b>271</b>
12.1	Overview of the implementation strategy . . . . .	271
12.1.1	Alignment model . . . . .	274
12.1.2	CVn type syllable alignment . . . . .	275
12.1.3	CV type syllable alignment . . . . .	275
12.1.4	Other syllable type alignment . . . . .	277
12.2	Waveform synthesis . . . . .	277
12.3	Illustrations of the synthesis process . . . . .	278
12.4	Stem-ML model for SY . . . . .	284
12.4.1	Design of the Stem-ML model . . . . .	284
12.4.2	Analysis of best-fit parameters of the intonation model . . . . .	285
12.5	Discussion . . . . .	287
12.6	Summary . . . . .	289
<b>V</b>	<b>Model Evaluation, Summary and Conclusion</b>	<b>290</b>
<b>13</b>	<b>Evaluation and discussion</b>	<b>291</b>
13.1	Experimental data preparation . . . . .	292



# CONTENTS

13.2	Quantitative evaluation . . . . .	293
13.3	Qualitative evaluation . . . . .	295
13.3.1	Intelligibility evaluation . . . . .	297
13.3.2	Naturalness evaluation . . . . .	299
13.4	Discussion . . . . .	301
13.5	Summary . . . . .	303
<b>14</b>	<b>Summary, conclusion and further work</b>	<b>305</b>
14.1	Summary . . . . .	306
14.2	Contribution to knowledge . . . . .	309
14.3	Extension of the model for other tone languages . . . . .	311
14.4	Future work . . . . .	312
<b>VI</b>	<b>Appendices</b>	<b>338</b>
<b>A</b>	<b>Inventory of 230 Yorùbá syllables</b>	<b>339</b>
<b>B</b>	<b>BNF for standard Yorùbá text</b>	<b>341</b>
B.1	Sample Yorùbá text composed for corpus data . . . . .	341
B.2	BNF for standard Yorùbá text . . . . .	342
<b>C</b>	<b>Design of the text markup system for TTS</b>	<b>343</b>
C.1	Introduction . . . . .	343
C.1.1	Text models . . . . .	344
C.1.2	Issues in SY typesetting and markup . . . . .	345
C.2	Text-to-Speech markup system . . . . .	345
C.2.1	Design of XML system . . . . .	346
C.3	Document Type Definition (DTD) . . . . .	347
C.3.1	The document tag . . . . .	347
C.3.2	The paragraph tag . . . . .	348
C.3.3	The sentence tag . . . . .	348
C.3.4	The phrase tag . . . . .	349
C.4	Text contents description tags . . . . .	350
C.4.1	The SAYAS tag . . . . .	350
C.5	Schema for the Document Type Definition (DTD) . . . . .	351
C.6	Markup of the first paragraph in the sample text . . . . .	353
<b>D</b>	<b>Sentences used for modelling</b>	<b>354</b>
D.1	Single phrase sentences . . . . .	354
D.2	Two phrase sentences . . . . .	356
<b>E</b>	<b>Labelled speech data</b>	<b>358</b>
<b>F</b>	<b>Program listings</b>	<b>363</b>
F.1	Program for extracting data from annotation speech files . . . . .	363
F.2	MatLab programs . . . . .	365
F.2.1	Stylisation interpolation program . . . . .	365

CONTENTS

F.3 *Praat* program listings . . . . . 369

# List of Figures

1.1	A graphical illustration of the relations between pitch tone and $f_0$ . . . .	19
1.2	A graphical illustration of prosodic space . . . . .	21
1.3	Research context . . . . .	29
2.1	Wheatstone’s reconstruction of von Kempelen’s machine (Source: <i>Lemmetty</i> (1999)) . . . . .	35
2.2	The Yorùbá talking drum . . . . .	36
2.3	The VODER speech synthesiser (Source: <i>Klatt</i> (1987)) . . . . .	38
2.4	Stages in modern TTS process . . . . .	40
2.5	Structure of the formant synthesiser used in <i>Lee et al.</i> (1993) . . . . .	49
3.1	SY syllable structures with examples (Note: $\phi$ implies Nil) . . . . .	60
3.2	Yorùbá syllable viewed as an object . . . . .	63
3.3	Sample scanned texts in SY orthography . . . . .	68
4.1	HMM based prosody model for TTS ( <i>Tokuda et al.</i> (2002b)) . . . . .	73
4.2	RNN based prosody model ( <i>Wang et al.</i> (2002)) . . . . .	77
4.3	Top line and Bottom line in the Pierrehumbert model . . . . .	79
4.4	Graphical illustration of the <i>Tilt</i> model . . . . .	83
4.5	The structure of Fujisaki model ( <i>Clark</i> (2003)) . . . . .	87
4.6	Block diagram of the Stem-ML algorithm ( <i>Kochanski and Shih</i> (2003)) . . . . .	91
4.7	Typologically distinct lexical tones automata ( <i>Gibbon</i> (2004a)) . . . . .	94
4.8	Gibbon architecture for prosody modelling ( <i>Gibbon</i> (2004a)) . . . . .	97
4.9	Implementation of Monaghan’s refinement of Ladd’s model ( <i>Monaghan</i> (2003)) . . . . .	99
6.1	Peaks and valleys on a waveform . . . . .	112
6.2	Illustration of S-Tree generation . . . . .	114
6.3	Example of fuzzy set . . . . .	118
6.4	Fuzzy logic Controller ( <i>Lee</i> (1990a)) . . . . .	118
7.1	SY TTS system overview . . . . .	132
7.2	Speech synthesis strategy . . . . .	136
7.3	Waveform and Spectrogram of SY word “Ita” (meaning “Outside”) . . . . .	138
7.4	Stages in our proposed prosody modelling . . . . .	141
7.5	Illustration of our prosody modelling technique . . . . .	143
7.6	Overview of our R-Tree generation technique . . . . .	144
7.7	Example waveform . . . . .	147
7.8	R-Tree of the waveform in Figure 7.7 . . . . .	148

## LIST OF FIGURES

8.1	Syllable distribution . . . . .	150
8.2	Linguistic unit distribution . . . . .	151
8.3	Wavesurfer Screen Capture . . . . .	154
8.4	Screen capture of annotation of the syllable <i>dé</i> . . . . .	157
8.5	Screen capture of annotation of the sentence <i>Ópé kí ó tó dé, kò tètè lọ</i> . . . . .	158
8.6	Example $f_0$ data for the sentence “ <i>Ópé kí ó tó dé, kò tètè lọ</i> ” . . . . .	159
9.1	Levels in Intonation modelling . . . . .	163
9.2	Domain of intonation phenomena in Yorùbá utterance . . . . .	164
9.3	A demonstration of IPO stylisation process ( <i>t Hart et al. (1990)</i> ) . . . . .	167
9.4	IPO standardisation from . . . . .	168
9.5	Early, medial and late peak alignment in KIM ( $V_{on}$ indicates the onset of the stressed vowel) ( <i>Braunschweiler (2003)</i> ) . . . . .	170
9.6	The quadratic monomial model (source: <i>Taylor (1994)</i> ) . . . . .	171
9.7	Calculation of a local target point in MOMEL (source: <i>Campbell (2000)</i> ) . . . . .	172
9.8	Raw $f_0$ of /CV/ syllable where $V=a$ . . . . .	176
9.9	Interpolation into $f_0$ curve of /Gba/ syllable . . . . .	179
9.10	Interpolation into $f_0$ curve of /Wa/ syllable . . . . .	180
9.11	Mean opinion score for naturalness . . . . .	185
9.12	Abstract curve of the three Yorùbá tones . . . . .	187
9.13	Signature for the standardised tones . . . . .	188
9.14	Graphical representation of co-articulated tones according to <i>Hombert (1976)</i> . . . . .	190
9.15	Main pseudocode for S-Tree generation . . . . .	193
9.16	Pseudocode for finding deepest valley . . . . .	194
9.17	Pseudocode for finding highest peak . . . . .	194
9.18	Transcript of S-Tree generation for the SY sentence “ <i>Wón sì tún pòwe Ìbàdàn.</i> ” . . . . .	196
9.19	Transcript of S-Tree generation for phrase “ <i>Bàbá àgbè tita kòkò</i> ” . . . . .	197
9.20	Transcript of S-Tree generation for phrase “ <i>kí ótó mòpé kòkò ti wón</i> ” . . . . .	198
9.21	Relational organisation of valleys in multi-phrase utterance . . . . .	199
9.22	Abstract intonation waveform for sentence “ <i>Bàbá àgbè tita kòkò, kí ótó mòpé kòkò ti wón</i> ” . . . . .	200
10.1	Natural and predicted $f_0$ peak plots for intonation models for the same tone sequence . . . . .	212
10.2	Computation of $f_0$ peak on a sentence containing syllables with the H, M, and L tone sequence . . . . .	213
10.3	Membership functions for the premise variables . . . . .	215
10.4	Relative Input-output behaviour as predicted by the model for different tone combination . . . . .	218
10.5	An algorithm for generating intonation contour for an utterance . . . . .	220
10.6	Computation of membership values for crisp input data . . . . .	222
10.7	Computed points on the transcript of S-Tree . . . . .	224
10.8	Natural versus Synthetic intonation of “ <i>Won si tun powe Ibadan</i> ” . . . . .	225
10.9	MOS results for the naturalness test . . . . .	232
11.1	Level 0 Factors on syllable duration . . . . .	242

## LIST OF FIGURES

11.2	Factors affecting syllable duration in SY utterance as observed in the training corpus . . . . .	245
11.3	Membership function of continuous duration affecting factors . . . . .	254
11.4	Membership function for the output . . . . .	254
11.5	FDT building algorithm . . . . .	257
11.6	FDT for numerical values duration factor . . . . .	258
11.7	Fuzzy decision tree for the duration model . . . . .	260
12.1	Structure of syllable data file . . . . .	273
12.2	Schematics of time and $f_0$ alignment . . . . .	276
12.3	The schematics of speech synthesis process . . . . .	278
12.4	Transcript of <i>S-Tree</i> generation for the sentence “ <i>Óní láti lọ wobè.</i> ” . . . . .	280
12.5	Transcript of <i>S-Tree</i> generation for the sentence “ <i>Òdòmi lódé, Kó tó lọ.</i> ” . . . . .	281
12.6	Result for the sentence “ <i>Óní láti lọ wobè.</i> ” . . . . .	283
12.7	Result for the sentence “ <i>Òdòmi lódé, Kó tó lọ.</i> ” . . . . .	283
12.8	Model fit and raw data for SY sentence “ <i>Óní láti lọ wobè.</i> ” . . . . .	287
12.9	Stem-ML prediction of $f_0$ and raw data for the two phrases of SY sentence “ <i>Òdòmi lódé, Kó tó lọ.</i> ”. This data is in the test set; the model prediction is based on parameters derived from the training set, using syllable boundaries for this specific utterance. . . . .	288
13.1	Result of intelligibility evaluation . . . . .	297
13.2	Result of naturalness evaluation . . . . .	300
13.3	G value evaluation . . . . .	301
C.1	Example SY Text . . . . .	346
E.1	SY syllable Bá (get to) . . . . .	358
E.2	SY syllable Dé (cover) . . . . .	359
E.3	SY syllable Gbẹ (to dig) . . . . .	359
E.4	Annotate file for the SY sentence “ <i>Bàbá àgbẹ ti ta kòkó.</i> ” . . . . .	360
E.5	Annotate file for the SY sentence “ <i>Ìyálójà ló mọ.</i> ” . . . . .	360
E.6	Annotate file for the SY sentence “ <i>Ó mọ pé èmi kọ.</i> ” . . . . .	361
E.7	Annotate file for the SY sentence “ <i>Ópẹ kí ó tó dé, kò tètè lọ.</i> ” . . . . .	361
E.8	Annotate file for the SY sentence “ <i>Ìyá tidè, ejé ká jẹun.</i> ” . . . . .	362
F.1	A screen capture of the interface to the <i>Praat</i> programs. . . . .	370

# List of Tables

2.1	Synthesis Units in some popular TTS . . . . .	53
3.1	Segmental phonemes of Standard Yorùbá consonant . . . . .	58
3.2	Segmental phonemes of Standard Yorùbá vowel . . . . .	58
3.3	Phonological structure of SY Syllable . . . . .	61
3.4	Statistics of Yorùbá syllables . . . . .	62
3.5	Example of syllable structure for the SY word <i>Ikán</i> . . . . .	65
3.6	Punctuation in SY text . . . . .	66
3.7	Other tokens in SY text . . . . .	66
3.8	A survey of text in Yorùbá textbooks and newspapers . . . . .	69
4.1	INTSINT coding system . . . . .	86
8.1	Comparison of syllables and tonal distribution in tone languages . . . . .	152
8.2	Speech database description . . . . .	153
8.3	Speech sound recording parameter . . . . .	155
8.4	Speech data based annotation symbols . . . . .	155
8.5	Duration data for a sentence . . . . .	159
9.1	Coefficient for linear interpolation of $f_0$ curve (Hz) / <i>Wa</i> / . . . . .	177
9.2	Coefficient for linear interpolation of $f_0$ curve (Hz) / <i>Gba</i> / . . . . .	177
9.3	Coefficient for quadratic interpolation of $f_0$ curve (Hz) for / <i>Wa</i> / . . . . .	177
9.4	Coefficient for quadratic interpolation of $f_0$ curve (Hz) for / <i>Gba</i> / . . . . .	178
9.5	RMSE (Hz) result for each interpolation for syllable / <i>Gba</i> / . . . . .	181
9.6	RMSE (Hz) result for each interpolation for syllable / <i>Wa</i> / . . . . .	182
9.7	Intelligibility of SY syllables with natural and synthetic $f_0$ curves . . . . .	184
9.8	Terracing rule in SY ( <i>Courtenay</i> (1971)) . . . . .	189
9.9	Phonological rules for SY tone interaction . . . . .	191
9.10	Depiction of intonation phenomenon in SY tone interaction . . . . .	191
10.1	A description of the fuzzy model variables and parameters in Equations 10.5 & 10.6 . . . . .	211
10.2	Linguistic label for <i>TonCon</i> . . . . .	213
10.3	Linguistic label for <i>RelPos</i> . . . . .	214
10.4	Detailed tone pattern for sentences shown in Figure 10.4 . . . . .	217
10.5	Minimum $f_0$ value perceptible as tone $T \in H, M, L$ a speaker . . . . .	219
10.6	Sentence data . . . . .	223
10.7	Premise and consequence computation for each rule . . . . .	226
10.8	Comparison of MSE results . . . . .	227

LIST OF TABLES

10.9	Comparison of quantitative results . . . . .	227
10.10	Comparison of intelligibility results . . . . .	230
10.11	Qualitative evaluation scores for naturalness . . . . .	231
10.12	Comparison of the naturalness test results . . . . .	233
11.1	Levels of syllable duration affecting factors . . . . .	239
11.2	Statistics for the characteristics of the training and test data sets . . . . .	240
11.3	Statistics of the duration of various SY syllable types . . . . .	243
11.4	Syllable duration affecting factors . . . . .	253
11.5	Syllable duration predicted . . . . .	253
11.6	A summary of our FDT variables, functions and parameters . . . . .	256
11.7	CART input Description file . . . . .	262
11.8	CART Tree for numerical duration affecting factors . . . . .	264
11.9	Optimal CART for duration model . . . . .	264
11.10	Result for quantitative evaluation for FDT . . . . .	266
11.11	Results for quantitative evaluation for CART . . . . .	266
11.12	Comparison of duration models (based on test set results) . . . . .	267
11.13	Results for the intelligibility evaluation . . . . .	267
11.14	Results for naturalness evaluation (Training set) . . . . .	269
11.15	Results for naturalness evaluation (Test set) . . . . .	269
12.1	Computed data for the sentence “ <i>Óní láti lọ wobè.</i> ” . . . . .	282
12.2	Computed data for the sentence “ <i>Òdòmi lódé, Kó tó lọ.</i> ” . . . . .	282
13.1	$f_0$ contour evaluation . . . . .	294
13.2	Results for intelligibility evaluation . . . . .	298
13.3	Qualitative evaluation scores . . . . .	299
13.4	Results for naturalness evaluation . . . . .	301
A.1	V and CV type syllable inventory <i>Total</i> = 133 . . . . .	339
A.2	Vn and CVn type syllable inventory <i>Total</i> = 95 . . . . .	340
A.3	<i>N</i> type syllable inventory <i>Total</i> = 2 . . . . .	340
C.1	L <sup>A</sup> T <sub>E</sub> X annotation for SY diacritic and under dots . . . . .	346
C.2	Tags for SY text . . . . .	350

# Part I

## Introduction



# Chapter 1

## Prosody modelling: a synopsis

A computer text-to-speech (TTS) synthesis system converts digital text into speech-like sounds. The potential applications of high-quality TTS systems are numerous. Some of these applications include: (1) language education, (2) reading aid to people with physical impairment, e.g. the visually-impaired, (3) interface to written materials for illiterate people, (4) test bed for studying and evaluating hypothesis in speech science and engineering, and (5) component in a voice-based Human-Computer Interface (HCI). Another beneficial application of high quality TTS systems is that it can be used for preserving endangered languages.

Modern TTS systems can logically be decomposed into four main modules: (i) text analysis module, (ii) linguistic structure generation module, (iii) prosody generation module and, (iv) speech signal generation module. Prosody modelling remains the most challenging problem in modern TTS research and development because contemporary synthetic speech still suffers from unexpressive and often inappropriate prosody (*Ogden et al.*, 2000). One reason for this is that the input to a TTS system contains little or no explicit information about how the prosody can be generated and such information is extremely hard to extract from plain text.

In addition, the extracted information, which are discrete, must be converted into continuous speech signal. This conversion process is not trivial because the complex interactions between different aspects of prosody are often poorly understood (*van Santen and Hirschberg*, 1994). For example, the translation of prosody phenomena, such as downstep or final lowering, which are perceptually significant, into precise

acoustic parameter is influenced by a large number of contextual factors which are yet to be fully understood (*Pitrelli et al.*, 1994; *Kohler*, 1997a; *Cahn*, 1998; *Goldsmith*, 1999; *Hirst et al.*, 2000; *Braunschweiler*, 2003; *Monaghan*, 2003). It has also been observed (*Ogden et al.*, 2000) that there is a fundamental lack of attention to the systematic fine details, which listeners expect to hear in human speech production.

A number of TTS systems have been developed for European languages, such as English and German (*Kohler*, 1986; *Möbius*, 2003; *Monaghan*, 2003), as well as Asian languages, such as Mandarin (*Shih and Sproat*, 1996; *Kochanski and Shih*, 2003) and Cantonese (*Lee et al.*, 2002a; *Li et al.*, 2004) Chinese. For these languages, the control of several prosodic parameters have been refined over many years and recent improvements have come from the resolution of theoretical details (*Monaghan*, 1990). However, work on TTS development for African languages is scarce. Some attempts have been made to develop TTS system for some languages spoken in Africa, such as Arabic (*El-Imama*, 1989, 1990) and Zulu (*Louw and Barnard*, 2002). Other attempts have been directed to modelling aspects of prosody for African languages, e.g. intonation modelling for Hausa (*Lindau*, 1986), Bamileke Dschang (*Bird*, 1994), and tone modelling for Tem and Baule (*Gibbon*, 2004a).

This thesis addresses an important research problem in the design and development of a text-to-speech synthesis system for the Standard Yorùbá language, i.e. the prosody modelling problem.

## 1.1 Definition of terms

Some terms used in this thesis can have different meaning in different fields of study. Since our research cuts across many fields, the following sections provide the definition of terms used in this thesis. These definitions are by no means formal but express the context in which we use the terms.

### 1.1.1 An utterance

In this thesis, an utterance is considered to be a stream of sounds generated as a result of the activities of the human speech organ or a device mimicking the speech organ.

We define natural utterances as those produced by the human speech organ. Other forms of utterance are artificial. The acoustic signal corresponding to an utterance, in the form of sound waveforms, creates a perceptual impression on the hearer. The perceptual impressions do not carry semantic and/or pragmatic connotations. An utterance is, therefore, made up of a stream of sounds which need not have any meaning. We express this symbolically as follows:

$$\textit{Utterance} = \textit{Stream of sounds}$$

### 1.1.2 Speech

This is an utterance with an underlying linguistic structure. If the linguistic structure underlying the utterance is a human language, then the speech is a natural speech, otherwise it is an artificial speech. To be heard as speech, time-varying acoustic properties must bear the right relationships to one another. A natural speech is a natural utterance plus a human (or natural) language. A perceived speech has a meaning (semantic and/or pragmatic) associated with it by the human listener. Spoken syllables, words, and phrases are portions of a speech. This definition can be expressed symbolically as:

$$\textit{Speech} = \textit{Utterance} + \textit{Language}$$

The plus operator, i.e. +, indicates integration of the two operands. In this context, nonsense words are utterance but not speech.

### 1.1.3 Text

A text is an abstract representation of speech. The representation can be in written form (e.g. hard-copy orthographic) or digital form, i.e. soft-copy on computer disk.

### 1.1.4 Fundamental frequency ( $f_0$ ), pitch, tone, and intonation

The *fundamental frequency* ( $f_0$ ) value is the numerical data representing a point on the fundamental frequency dimension of a speech signal. We define a sequence of  $f_0$

values over the speech waveform of a syllable as the  $f_0$  curve on that syllable. The  $f_0$  curve is distinguished by a peak and/or a valley which corresponds to the turning points in the curve. We define a sequence of  $f_0$  values over the speech waveform of a linguistic unit longer than a syllable, i.e. word, phrase, or sentence, as an  $f_0$  contour. Simply stated, an  $f_0$  contour comprises more than one  $f_0$  curves.

*Pitch* is the perceptual correlate of the fundamental frequency ( $f_0$ ) curve. We regard pitch as a subjective abstract property of speech sound because of the discrepancies between  $f_0$  pattern and the perceived pitch. A *tone* is the symbolic representation of the pitch corresponding to the  $f_0$  curve. *Intonation* is the symbolic representation of the pitch pattern corresponding to an  $f_0$  contour.

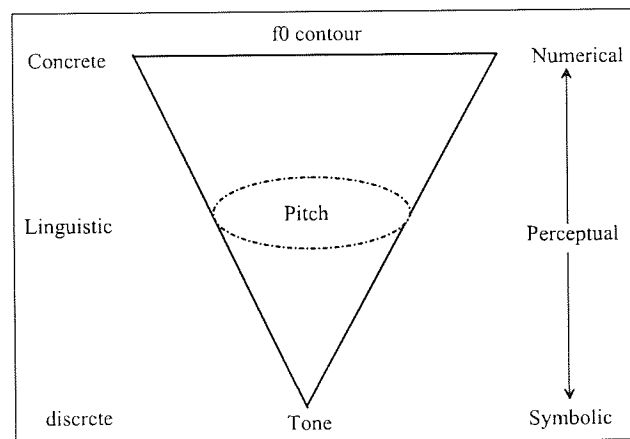


Figure 1.1: A graphical illustration of the relations between pitch tone and  $f_0$

It is important to note that the four terms defined here refer to different aspects of the same entity. What people perceive when they hear a sound is a pitch. The acoustic correlate of what is perceived is the fundamental frequency ( $f_0$ ). The symbolic representation of the  $f_0$  is a tone when it applies to a syllable or an intonation when it applies to units longer than a syllable e.g. a word.

A range of  $f_0$  values can correspond to the same pitch and different types of pitch can be represented using the same tonal signature (see Figure 1.1). For example, in our data,  $f_0$  curve with a peak value of  $120Hz$ ,  $130Hz$ , or  $150Hz$  is perceived as a high tone for an adult male speaker. Similar  $f_0$  curves with the same peak values are perceived as a mid tone for a female speaker.

We emphasise here that  $f_0$  is a measurable quantity with a numerical value and

pitch is a linguistic entity that can only be represented qualitatively whereas tone and intonation are abstract entity which can only be assigned discrete symbolic labels. By this definition, intonation is regarded as an entity that can only be modelled symbolically whereas  $f_0$  contour is a numerical entity that has continuous values and can be represented using mathematic functions.

One of the tenets of this work is that syllable based pitch contours are the real focus in Yorùbá, rather than intonation in the sense in which it is applied in non-tone language such as English. Hence, throughout this thesis, we use the term ‘intonation’ to mean “syllable based pitch contour”.

### 1.1.5 Duration

*Duration* is the span in time occupied by a linguistic entity, e.g. syllable, in a speech waveform. The time tier or axis of a speech waveform is, therefore, regarded as its ordering parameter. We assume that the durational magnitude of a linguistic entity is directly proportion to its perceptual length. This means that a speech waveform that occupies wider span on the time axis will be perceived as having a longer duration.

### 1.1.6 Prosody

From the physical point of view, speech *prosody* can, putatively, be defined as having three dimensions: (i) fundamental frequency ( $f_0$ ) contour, measured in Hertz ( $Hz$ ); (ii) duration, measured in milliseconds ( $msec$ ); and (iii) intensity, measured in decibel ( $dB$ ). These dimensions define the *prosodic state-space* in an acoustic signal (see Figure 1.2). These physical features constitute the prosody of language and provide important information about meaning and structure (*Fernanda, 2000*).

The prosodic state-space has a perceptual correlate which can be described symbolically using phonological theories. Prosody gives speech sound unique perceptual impressions and communication characteristics. Speech prosody can also be characterised by linguistic and para-linguistic features. The linguistic features include the structure of the spoken language and how linguistic units such as syllables, words, phrases and sentences are organised in a speech. Para-linguistic features include cognitive attributes of speech, such as gender, age, emotions, etc., which are speaker and

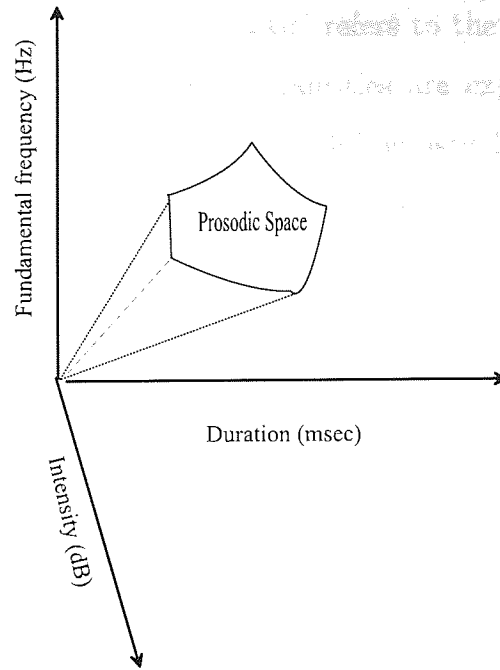


Figure 1.2: A graphical illustration of prosodic space

language dependent.

It is well known that a wide range of acoustic features contribute to the perceptual coherence in prosody. The influence of some, such as the fundamental frequency, timing and intensity is widely acknowledged (*Bradlow et al.*, 1996). Other features, such as speaker characteristics, utterance mode and mood, are known to be important but their contributions to prosody is yet to be understood. It is also apparent that listeners use various acoustic cues to identify naturally-produced prosody. When such cues are absent in synthetic speech, it reduces the intelligibility and naturalness of such speech. Modelling fine phonetic details of these acoustic cues is important since their variations contribute to making the time-varying speech signal an effective communicative medium (*Ogden et al.*, 2000; *Braunschweiler*, 2003).

### 1.1.7 Linguistic

The term ‘linguistic’ is used in two different contexts in this thesis. In the first context, we use it to imply the scientific study of language, covering the structure (morphology and syntax), sounds (phonology), and meaning (semantics) as well as sociolinguistics, and pragmatics.

In the second context, the term ‘linguistic’ refers to the label used for identifying a qualitative class in fuzzy logic. Fuzzy quantities are expressed in terms of fuzzy numbers, or fuzzy sets, which are associated with linguistic labels (*Lee, 1990a; Sugeno and Yasukawa, 1993*). Membership functions are used to convert numerical values into fuzzy linguistic values. This function determines the degree to which a number belongs to a fuzzy set. The membership function is usually formulated to assign values in the closed interval  $[0.0, 1.0]$ , depending on the degree to which a numerical value belongs to a linguistic class. The value 0.0 is for non-membership and 1.0 for full membership. For example, the  $f_0$  curves with peak values of  $120.00Hz$ ,  $150.00Hz$  and  $180.00Hz$  can be said to belong to the fuzzy linguistic class “High” tone to a degree 0.7, 0.8, and 0.9 respectively.

Throughout this thesis, we use the first definition of linguistic as the default. The second definition is used in the description of our fuzzy logic models.

## 1.2 Issues in prosody modelling

The ultimate goal in text-to-speech synthesis technology is to develop an ideal TTS system. The synthetic speech produced by such a system should be indistinguishable from that of a human reading a piece of text aloud. Text-to-speech synthesis by computers has reached a high level of performance, with increasingly sophisticated models of linguistic structure, low error rates in text analysis, and high intelligibility. However, the quality of the synthesised speech is still not good enough for the ubiquitous tasks required of such technology in modern applications (*Lieberman, 1995*). This poor quality has been attributed to the inability of modern TTS systems to adequately model speech prosody at an acceptable level of accuracy. Therefore, the development of a prosody system that can deliver high quality synthetic speech remains one of the major problems being addressed by speech scientists and engineers.

The relative ease with which human performs the speech production task may suggest that speech synthesis is a simple or trivial task. For example, in a normal speech, a speaker produces approximately between three to six syllables per-second (*Wang and Chen, 1994*). This speech is produced with a remarkably high accuracy, e.g. less than

one error per thousand words (*Costa et al.*, 2000). However, the mechanism involved in speech production and the relationship between this mechanism and speech prosody are very complex.

The confounded nature of the various dimensions of speech prosody makes it difficult to separate the different kind of information required for modelling and realising it. For example, it is difficult to separate the knowledge for encoding the phonology and those for encoding the phonetics of prosody when describing its manifestation in acoustic signal. The transition from the phonological through the phonetics to the acoustic tiers of prosody is a fuzzy and ambiguous one. When these knowledge are developed for synthesis in separate modules, the problem of integrating them becomes more difficult (*Ogden et al.*, 2000).

In addition, the extraordinary aspect of speech waveform is that it does not only encode the communicated ideas and concepts, but it also encodes the unique characteristics of the speaker. The encoded characteristics of the speaker cover a wide spectrum of behavioural patterns. At one end of the spectrum are the more stable characteristics such as speakers' voice, which is determined by the physiological structure of the speech organ. At the other extreme is the dynamic aspect of speaker's characteristics, such as the speaker's emotions, which are determined by the state of mind at the point of speaking. Other factors that might affect spoken utterance include the source of spoken item (e.g read or spontaneous). The tools provided by the current state of the art in speech technology cannot analyse this behavioural pattern in a definitive manner.

Moreover, written text does not contain sufficient cues for determining the prosody of the corresponding speech and the current state of knowledge on how prosody information is organised to generate the overall rhythmic and melodic aspect of natural speech is rather inadequate (*Huckvale*, 1997, 2002).

### 1.3 Motivation

While TTS for European and Asian languages has received a considerable amount of research attention, work on African languages is rare. One of the motivations of



this work is to contribute to knowledge in this area of study by experimenting with an African language, i.e. Standard Yorùbá (SY). There are a few work on the speech science of SY and most have proposed interesting but inconclusive phonological and phonetic theories about the prosody of SY. To verify such theories, a practical model is required to serve as a test-bed for evaluation purpose. An important requirement of such model is flexibility and modularity. By this, we mean that the model should facilitate the independent implementation of various dimensions of speech prosody so that we can test different models and select the best based on their performances. In the following subsections, we will discuss the theoretical and practical motivations of the approach adopted in this research.

### 1.3.1 Theoretical motivation

Prosody in speech synthesis have been developed around two major classes of intonation phonology models: Autosegmental-Metrical (AM) model (*Lieberman and Pierrehumbert, 1984; Ladd, 1996*) and Hierarchically Organised (HO) model (*Fujisaki and Hirose, 1982; Grønnum, 1992*). The AM models cover a group of theories that take the view that intonation is made up of a series of discrete pitch movements. The HO models, on the other hand, interpret intonation as a complex pattern resulting from the superposition of several components. The superposition model also argues that there are strong interactions among the different layers of the intonation hierarchy. The hallmark of intonation realisation in the AM approach are its strict locality and its temporal asymmetry. This is contrast to HO models in which the utterance  $f_0$  contour arises through superposition of its component phrasal  $f_0$  contour with local accent-related  $f_0$  curves.

Inherent in the HO model is the idea that speakers pre-plan intonation a few syllables ahead during speech production. This idea is rejected by the AM approach because it is considered implausible as pre-planing will increase the cognitive load of speech production. It has been shown that the AM approach works well for non-tone languages such as English (*Pierrehumbert, 1981*). Findings in tone languages, however, indicate that its application is unlikely to produce an accurate model of intonation. For instance *Xu (1999a)* showed that intonation phenomena such as downstep, antici-

patory raising and carryover lowering contribute separately to the overall downtrend in Mandarin intonation.

The Stem-ML model *Lee et al. (2002a)*; *Kochanski et al. (2003b,a)* is another successful prosody model for tone language which explicitly incorporate parameters to model pre-planning in intonation production.

Two findings by Láníran and Clements suggest that the principle of pre-planning may indeed operate in SY intonation production. In *Láníran and Clements (1995)*, they showed that high tone raises in anticipation of downstep. In another work (*Láníran and Clements, 2003*), their experimental results showed that some SY speakers use small but significant pre-planning in tone production, although the results do not generalise to all speakers. Other works (*Connell and Ladd, 1990*; *Láníran and Clements, 1995*), have shown that SY intonation production exhibits intonation phenomena such as H raising, L lowering and final lowering, in which the production of a target tone requires information about the preceding and following tones. All these findings point to the fact that some elements of pre-planning are employed during the production of SY intonation.

Furthermore, *Silverman and Pierrehumbert's (1990)* findings on tonal alignment suggest the need for looking ahead to the upcoming tones when producing the current tone. Therefore, the  $f_0$  scaling employed by the AM approach may not be successful in adequately describing the intonation in SY prosody. This is partly because the tonal signature that counts perceptually and phonologically may, from a mathematical point of view, not have a well-defined  $f_0$  curve (*Pierrehumbert, 2000*). This finding is particularly important for tone languages such as SY and Mandarin because tones are closely associated with syllables and are normally used to distinguish the lexical class of words that are made up of the same phonetic syllables. Variations in fundamental frequency ( $f_0$ ) during the production of a syllable are central to the accurate perception of the syllable tone as well as utterance intonation.

In the context of our TTS synthesis system development, there is the need to harmonise these approaches and apply them at the appropriate level of system design. For example, the phonological information in text, i.e. tones, are discrete and sequential. However, the intonation patterns generated from them are continuous and hierarchical

(*Collier*, 1990). Therefore, a unified framework is required to implement intonation and prosody within this context. A powerful tool in metrical phonology, i.e. the tree data structure, provides a starting point for the development of such framework. A tree-based approach incorporates both linear and hierarchical elements which can be exploited at various level of our intonation modelling. This motivated the use of tree-based structures in our prosody model.

### 1.3.2 Practical motivation

The next issue is how to realise prosody from the abstract phonological structure discussed above. Since the physical realisation of “prosody” can be observed and quantified from acoustic speech signal using  $f_0$ , duration and intensity, the ultimate aim in our TTS synthesis is to reproduce these three dimensions of speech prosody as accurately as possible by implementing a model that predicts, from an input text, all prosody phenomena that are known to be perceptually significant in each of the identified dimensions.

In the past, the prosody modelling problem has been viewed in terms of components which can be modelled individually. For example, there are several models in which intonation, duration and sometimes intensity, are viewed as independent entities. They are then modelled within separate framework and combined afterward. We view all dimensions of prosody as supplementing each other and that their modelling is best done in a unified framework.

Our approach to prosody modelling is based on the idea that the abstract (i.e. linguistic) and physical (i.e. acoustics) forms of prosody can be modelled within a modular unified framework which allows them to be viewed holistically while at the same time facilitating their independent implementation. To implement this idea, we used a combination of tree-based algorithms and fuzzy logic techniques to develop computational model of prosody.

The acoustic signal is computed using fuzzy logic (*Zadeh*, 1972) based rules which are constrained by perceptual information. These parameters are determined by fitting a fuzzy logic (*Zadeh*, 1972; *Sugeno and Yasukawa*, 1993) based model into a speaker’s speech data set. This enables our model to predict  $f_0$  values according to the degree

of evidence available in the data. This is achieved by using a linguistically motivated technique, i.e. fuzzy logic, to generate  $f_0$  and duration data in our prosody model. This approach increases the simplicity and predictive power of the model as well as maintaining the linguistic meaningfulness of the model parameters.

Our choice of the fuzzy logic approach is partly motivated by *Taylor's* (2000) argument that the descriptions of intonation phenomena are somewhat analogous to how people describe physical entities such as temperature. People perceive temperature as a continuum with no distinct categories. It is, however, helpful to have terms such as *hot* and *cold* to describe certain temperature situations. *Taylor* (2000b) argues that while there will be a lot of agreement as to what constitutes a hot or cold temperature under certain conditions, there will always be temperatures between these two extremes which could be described as either.

In the same manner, if we take the  $f_0$  range as a continuous variable measure in Hertz, the  $f_0$  equivalent of the pitch perceived for a typical H tone will be different from that of a typical M tone. However, there is a range between these  $f_0$  values that are ambiguous and can be categorised as either H or M based on the  $f_0$  values. Since our goal has been to describe observable phenomena and represent them as accurately as possible, in order to generate acceptable perceptual sensation in the synthetic prosody, we need to account for these ambiguities. In our opinion, the fuzzy logic technique, which has proven to be successful in modelling control systems for physical signals with no distinct categories (*Takagi and Sugeno*, 1985), is a viable approach for realising intonation contour.

Furthermore, *Demichelis et al.* (1983), *Lin et al.* (2003) and *Wang and Qiu* (2003) have shown that the linguistic concepts used for describing intonation phenomena are better modelled using the approximate reasoning approach. In addition, fuzzy logic can express the uncertainty and imprecision that are prevalent in intonation description. Fuzzy logic allows a number of alternative intonation patterns and associated parameters to be managed in a coherent manner. This attribute of fuzzy logic is particularly useful in selecting the best alternative amongst a set of possible conflicting intonation system designs.

Moreover, the simple structure of fuzzy rules makes it easier to include meaningful

linguistic terms which linguists and phoneticians are familiar with. This way, the model becomes easier to validate, evaluate, and improve. The simplicity of fuzzy rules and tree-based models has the advantage that the resulting prosody model can be easily understood, improved and also transported from one speaker or situation to another. The prosody model can also be valuable from a linguistic point of view, as many of the model parameters can be directly used to answer linguistic questions.

Our approach is motivated by the need to create a robust framework within which the qualitative and quantitative aspects of speech can be defined and accurately represented to the extent that the conversion of text into speech can be realised computationally.

## 1.4 Focus and scope of research

### 1.4.1 Research context

Our prosody model is intended as a part of an engineered software for speech synthesis. Its design is approached in the context of intelligent system engineering. Within this context, the development of a system involves the synthesis of knowledge from various sources using artificial intelligence techniques and tools. The knowledge that is used in the design of our model cuts across various fields of study. This includes language engineering, computer science, phonetics, linguistics and artificial intelligence. Generally speaking, this knowledge can originate from three main fields of study (see Figure 1.3): (i) Speech science and engineering, (ii) Artificial intelligence and (iii) Computer science and engineering.

The design and implementation of our model is governed by what can be practically achieved computationally. The evaluation of our model is constrained by native speakers' perception of the speech quality.

### 1.4.2 Research scope

This work focuses on prosody modelling in the context of tone language TTS applications. The SY language is used to demonstrate the applicability of our model. The linguistic and acoustic aspects of speech as well as perceptual information obtained

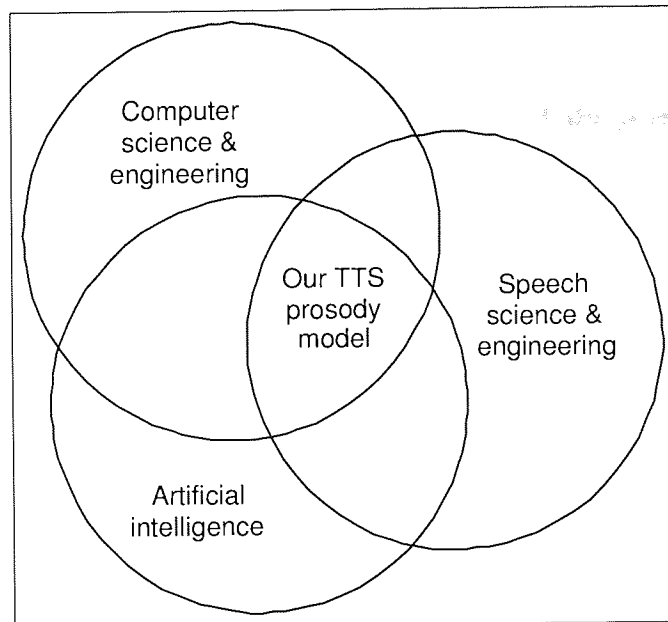


Figure 1.3: Research context

from native speakers' evaluation of natural and synthetic speech will be utilised in the design, implementation and evaluation of our model.

In the context of our prosody modelling, para-linguistic features of speech prosody, such as speaker's emotions and speaking styles are not explicitly considered. This is because they are too unstable to allow for practical computational analysis.

Our model involves the integration of the three dimensions of speech prosody: fundamental frequency, duration and intensity. In the present work, however, we have implemented the fundamental frequency and duration dimensions only. The intensity dimension has not been implemented due to the lack of literature on SY intensity as well as the time limit for this project.

We have used statement sentences to develop our prosody model. This decision is motivated by the findings by *Yuan et al.* (2003) for Mandarin, *Connell and Ladd* (1990) for Yorùbá and *Jun and Oh* (1996) for Korean, that the mode of a sentence, i.e. statement, question, exclamation, etc., does not have a significant effect on intonation pattern in tone languages. We assume that the components of a spoken sentence, i.e. syllable and word, co-articulate. We, however, ignore "inter-sentence" co-articulation.

### 1.4.3 Empirical hypothesis

We intend that our prosody model serves a number of purposes: (i) a model and framework for implementing prosody component for tone language TTS system, (ii) a test bed which facilitates the testing of various intonation and prosody theories. The main hypothesis is that prosody cues can be directly mapped onto phonological units of syllable. To put this theory into context, some underlying assumptions of the modelling approach are:

- At the abstract level, prosodic entity can be visualised and conceptualised as an holistic phenomena.
- At the physical level, prosody dimension can be independently realised and implemented within the structure specified by the abstract level.
- An intermediate level can be defined which facilitates the realisation of a computational relation between the abstract and physical levels.

We utilised the R-Tree technique to implement the holistic view of prosody. Within the structure defined by the R-Tree, we implemented the realisation of the fundamental frequency dimension using fuzzy control rule. The duration dimension is implemented by experimenting with the fuzzy decision tree and the classification and regression tree techniques.

## 1.5 Yorùbá: a brief introduction

Yorùbá is one of the major languages spoken in Africa. Other languages in this category include Arabic, Fulfude, Hausa, Lingala, Swahili and Zulu. Yorùbá has a speaker population of more than 30 million in West Africa alone (*Taylor, 2000a*). It has many dialects, but all speakers can communicate effectively using Standard Yorùbá (SY). SY is used in language education, mass media and everyday communication. The present study is based on the SY language.

SY is a tone language having three phonologically contrastive tones: High (H), Mid (M) and Low (L). Phonetically, however, there are two additional allotones or tone variant namely, rising (R) and falling (F) (*Bámgbóṣé, 1966; Connell and Ladd, 1990; Láníran and Clements, 2003*). The SY alphabet has 25 letters which is made up of 18

consonants and seven vowels. There are five nasalised vowels in the language and two pure syllabic nasals (*Bámgbóşé*, 1965; *Adéwólé*, 1988).

SY has a well-established orthography which has been in use for about ten decades. SY is relatively well studied when compare with other African languages and there are literature on the phonetics and phonology of the language. The present work is the first to examine prosody modelling for SY in the context of computer text-to-speech synthesis technology. A more detailed description of SY is provided in Chapter 3.

In this thesis, all the Yorùbá orthographic sentences and phrases appearing in the text and those appearing in the appendixes are translated in English. The SY orthographic materials appearing in the text are also glossed word-by-word in English. The word-by-word gloss is enclosed in square brackets, i.e. [ ], and are written before the literal English translation.

## 1.6 Thesis structure

The rest of this thesis is divided into four main parts and an Appendix. Their organisation is as follows.

Part II contains four chapters which describe the research background and provides a comprehensive literature review on modern prosody modelling approaches used in speech synthesis. In Chapter 2, the background of speech synthesis technology is presented. The Standard Yorùbá language is described in greater details in Chapter 3. Chapter 4 presents a review of literature on intonation modelling while Chapter 5 presents a review of literature on duration modelling.

Part III contains three chapters which describe the design methodology employed in our prosody modelling. In Chapter 6, the tools and techniques used in the design of our prosody model are presented. The model conceptualisation and design is presented in Chapter 7. Chapter 8 describes the data used in implementing our prosody model.

Part IV contains four chapters which discuss the implementation of our prosody model. In Chapter 9, the design of the prosody model, as it applies to SY, is presented. Chapter 10 and 11 contains the intonation and duration modelling respectively. The implementation of the complete prosody model is presented in Chapter 12.



Part V of the thesis contains two chapters. In Chapter 13, the model evaluation is presented while Chapter 14 concludes the present work.

Part VI contains the appendices which highlight some important information and discussion relevant to the work presented in this thesis.

## Part II

# Research Background and Literature Review

## Chapter 2

# Background of speech synthesis technology

### 2.1 A brief history of speech synthesis technology

Although the origin of speech and language is shrouded in mystery, the variety of languages and their associated sound patterns as well as the complexities involved in speech generation have been an interesting area of study since antiquity. As the human technology evolves from pure manual through mechanical, electrical, electronic and now artificial intelligence, so also does speech synthesis technology. In the following subsections, we discuss a brief overview of the evolution of speech synthesis technology. More comprehensive reviews can be found in published materials, from which much of the current material is drawn (*Klatt*, 1987; *Allen*, 1992; *Schroeder*, 1993; *Lemmetty*, 1999).

#### 2.1.1 The manual and mechanical era

The earliest speech synthesis systems are manual systems, fabricated from carved wood and animal skin. An important characteristic of these systems is that they are constructed in the form of musical instruments in that they consist of mechanisms for generating sound and a means for controlling the pattern of sound.

A well documented contribution in this era is an instrument developed by Christian Kratzenstein in 1779 (*Klatt*, 1987). It consists of pipes which act as acoustic resonators,

similar to the human vocal tract. These pipes can be activated by vibrating reeds. This instrument can be configured to produce the five Russian long vowels: (/a/, /e/, /i/, /o/, and /u/). For example, the sound /i/ is produced by blowing into the lower pipe without a reed causing the generation of the flute-like sound (Lemmetty, 1999).

Another important contribution in the mechanical era is the 'Acoustic-Mechanical Speech Machine' developed by Wolfgang von Kempelen in 1791 (Schroeder, 1993; Lemmetty, 1999). The essential parts of the machine were a pressure chamber which models the lungs, a vibrating reed to model the vocal cords, and a leather tube to model the vocal tract action. It is possible to produce different vowel sounds by manipulating the shape of the leather tube. Consonants were simulated by four separate constricted passages and controlled by the fingers. For plosive sounds, a model of a vocal tract that included a hinged tongue and movable lips were employed. von Kempelen's machine was able to produce single sound and combination of sounds in French and Italian. An advanced version of von Kempelen's speaking machine was constructed by Charles Wheatstone in about mid 1800's (see Figure 2.1).

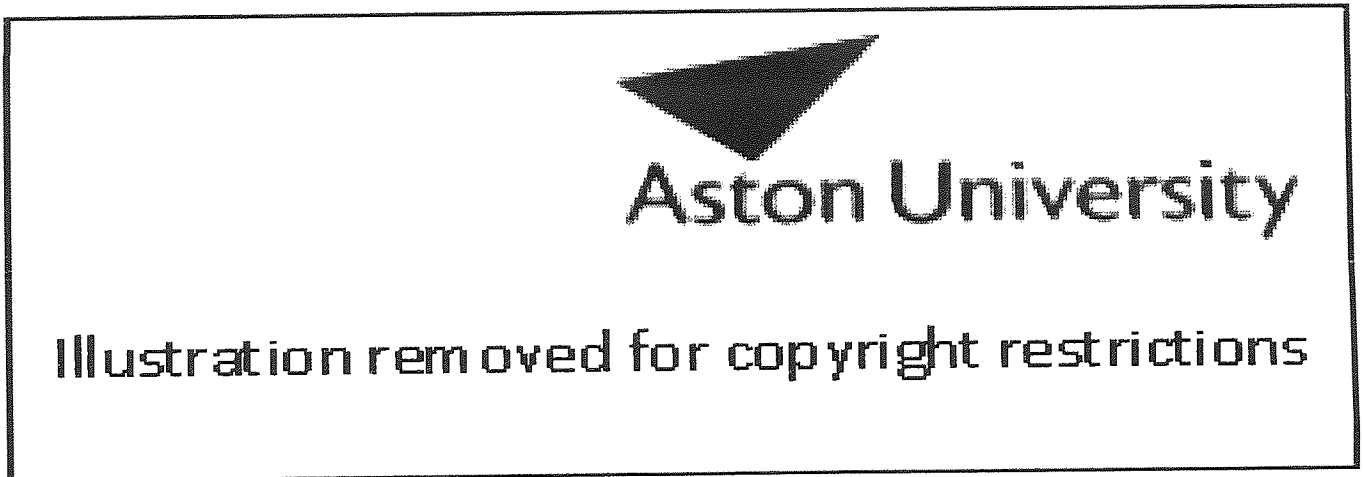


Figure 2.1: Wheatstone's reconstruction of von Kempelen's machine (Source: Lemmetty (1999))

The idea that human-like speech can be generated by mechanical means has been part of the Yorùbá culture since antiquity. The Yorùbá talking drum (*Ìyá ìlù*) is a crude, but perhaps, an important contribution which indicates that interest in speech synthesis technology transcends cultural, language and racial boundaries.

To make the talking drum, specially processed animal skins are used to cover the open ends of a wooden frame caved in the form of an hourglass. The wooden frame is about three feet long and the diameters of each of the open ends are about one foot. The skins are held into position over the openings of the wooden frame via a number of tightly drawn leather cords (see Figure 2.2). The leather cords are also made from animal skin.

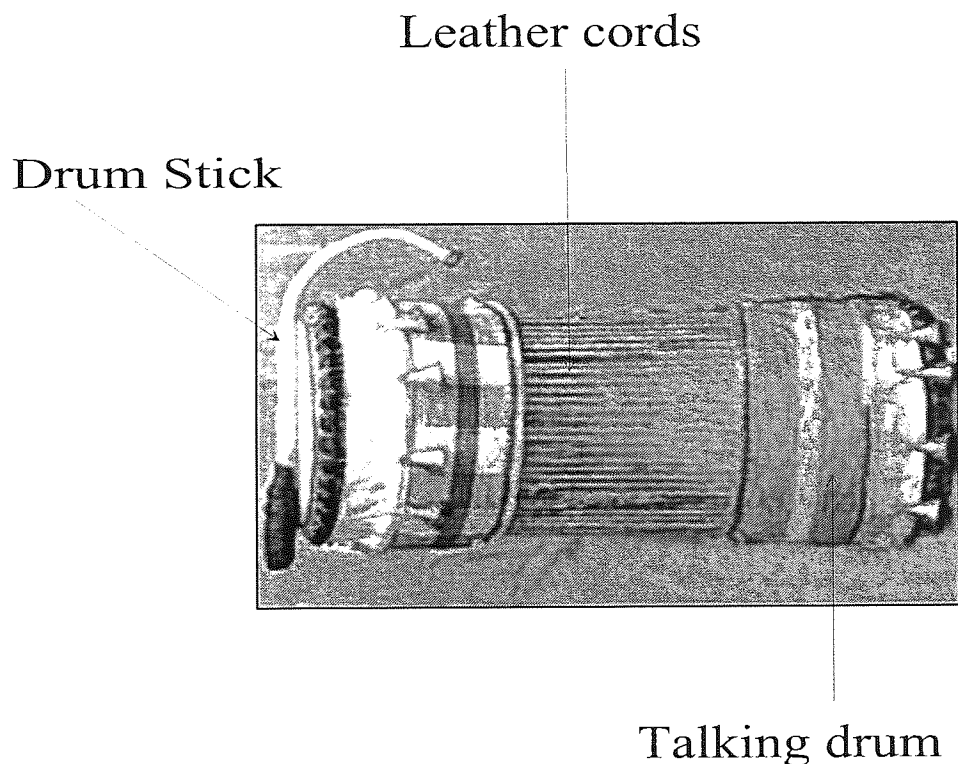


Figure 2.2: The Yorùbá talking drum

A considerable range of pitch can be generated by striking the animal skin cover using a curved drumstick, which is approximately 15 inches long. Held under the arm, the pitch of the generated sound can be varied by squeezing the leather cords. To create nasal-like sounds, the skin covering the hourglass-shaped wooden frame is manipulated with the finger nails.

Since Yorùbá is a tone language, the drum is normally manipulated to generate the pitch pattern corresponding to the desired utterance. Listeners can easily follow the message such as praise, insult, warning, song or proverbial saying, in the sound pattern. However, the ability to follow the message requires some familiarity with the context of the drumming.

Like the von Wolfgang's machine, the talking drum contains air chamber(s) made from animal skin. In both cases, the manual manipulation of this skin resulted in the generation of sound that are assumed to be human-like. Whereas, the von Wolfgang machine uses lever and reeds for manipulating the skin and its vibration, a drum stick and the finger nails are used in the case of the talking drum.

### 2.1.2 Electrical and electronic era

A major contribution to speech synthesis technology during the electrical era was the first full electrical speech synthesis device introduced by Stewart in 1922 (*Klatt, 1987*). The synthesiser consists of a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. A similar device by Wagner (*Flanagan, 1972; Flanagan and Ishizaka, 1976*) consisted of four electrical resonators connected in parallel and it was excited by a buzz-like source. The outputs of the four resonators were combined in the proper amplitudes to produce vowel spectra. The first three formants are generally considered to be enough for generating intelligible synthetic speech.

The development of the spectrogram representation of speech and the *source-filter* model of speech production are a major contribution that impacts on the rapid advancement of the speech synthesis technology. They facilitate the creation of parametric models of speech signals and the control of the models using rules. One of the first devices to exploit this idea was Voice Operating Demonstrator (VODER) introduced by Homer Dudley in 1939 (*Flanagan, 1972; Klatt, 1987*). The VODER consisted of a wrist bar for selecting a voicing or noise source and a foot pedal to control the fundamental frequency (see Figure 2.3). The source signal was routed through ten band-pass filters whose output levels were controlled by fingers.

The electronic era, which started around 1920, witnessed a number of improvements



Figure 2.3: The VODER speech synthesiser (Source: *Klatt* (1987))

in speech synthesis systems. Electronic technology provided the necessary framework for the development of a sophisticated speech synthesis system based on a model of the human articulatory mechanism. The first formant synthesiser during this era was the Parametric Artificial Talker (PAT), which was introduced by Walter Lawrence in 1953 (*Klatt*, 1987). PAT consisted of three electronic formant resonators connected in parallel. The input signal can come from two sources, namely; buzz or noise. A moving glass slide was used to convert painted patterns into six time functions to control the three formant frequencies, voicing amplitude, fundamental frequency, and noise amplitude.

Later in 1958, George Rosen introduced the first articulatory synthesiser called Dynamic Analog of the VOcal tract (DAVO) (*Klatt*, 1987). DAVO was controlled by tape recording of control signals created by hand. In mid 1960s, first experiments with Linear Predictive Coding (LPC) were made (*Schroeder*, 1993). Linear prediction was first used in low-cost commercial systems, such as TI Speak'n'Spell in 1980 (*Klatt*, 1987).

## 2.2 State-of-the-art in text-to-speech synthesis

The speech synthesis problem is only one half of the problem of text-to-speech synthesis. The speech that is synthesised by a text-to-speech synthesiser corresponds to the contents of an input text. The speech synthesiser must, therefore, be able to figure out an appropriate sequence of sounds as well as assign correct prosody to the input text before the speech synthesis can begin.

In speech synthesis, the state-of-the-art is a moving target. However, the fundamental techniques used as the systems evolve are quite similar. Modern TTS systems evolve from the results of research between late 1960 and early 1980. A product of that era is the first full text-to-speech system for English which was developed in 1968 by Noriko Umeda and her colleagues (*Klatt*, 1987). This machine was based on an articulatory model and included a syntactic analysis module with sophisticated heuristics. In 1979, Allen, Hunnicutt, and Klatt developed the MIT laboratory text-to-speech system called MITalk (*Allen et al.*, 1987) and two years later Dennis Klatt introduced his famous KlatTalk system (*Klatt*, 1987).

The technology used in the MITalk and KlatTalk systems form the basis for many modern synthesis systems, e.g. DECTalk and Prose-2000 (*Klatt*, 1987; *Allen*, 1994). The intelligibility of this technology is better than those in the electronic era but the speech generated still sounds robotic making it less pleasant to listen to. Modern speech synthesis systems employ advanced techniques in Speech Signal Process, Speech Science, Computer Science and Artificial Intelligence. In the following subsection, we review the main published works on modern TTS synthesis systems.

### 2.2.1 Modern TTS systems

The modern TTS system converts text into “synthetic speech” sound in a two-stage process (*Klatt*, 1976) (see Figure 2.4). The first stage, i.e. High Level Synthesis (HLS), reads the input text and generates a representation of how the text will be pronounced. The HLS stage is implemented using two modules. The first module, i.e. text-analysis module, analyses the input text to identify its basic elements and the context in which they are used. The results of the text-analysis module is fed into the second module,



i.e. prosody module, which generates a linguistic description of how the text will be pronounced. It also integrates timing and rhyme information into the generated representation. All the processing involved in this stage are together called High Level Synthesis (HLS) and the technology for implementing them is drawn from the domain of Natural Language Processing (NLP) and computational linguistics (*Sproat et al.*, 1996; *Shih and Sproat*, 1996).

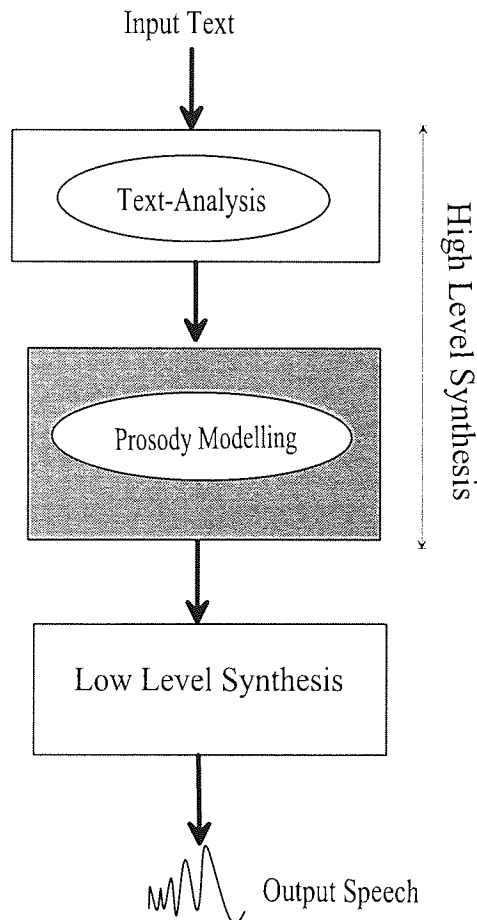


Figure 2.4: Stages in modern TTS process

The second stage, called Low Level Synthesis (LLS), takes the linguistic description outputted from the HLS stage as input and generates the corresponding speech waveform. The ultimate goal of this stage is to generate speech signal which, as much as possible, mimics the acoustic behaviour of speech produced by a native speaker reading the text aloud. There are three methods used to realise the LLS modules in the literature. These are Articulatory synthesis method (*Rubin et al.*, 1981; *Parthasarathy and Coker*, 1992; *Rahim et al.*, 1993; *Wouters*, 2001; *Beskow*, 2004), Formant synthesis

method (*Raptis and Carayannis, 1997; Jitca et al., 2002*) and Concatenative synthesis method (*O'Brien and Monaghan, 2001; Wu and Chen, 2001; Lee and Cox, 2002*).

In the following subsections, we will provide a brief review of literature on the HLS and LLS synthesis modules of modern TTS system with an emphasis on tone-language TTS.

### 2.2.2 Review of High Level Synthesis (HLS)

The output of the HLS stage is crucial to the production of accurate speech waveform. The HLS task comprises of text analysis and prosody generation. We review works on text analysis in this section. A detailed review of prosody modelling literature is presented in Chapters 4 & 5.

The text-analysis problem is language dependent, since different language uses different systems of orthography and expression. Therefore, the design of the text analysis module is a major focus in many TTS system for a new language. Text segmentation and normalisation are basic to all TTS systems but more complicated processing such as word end detection, syntactic analysis, morphological analyses and syllabification may be incorporated to account for language specific features or for the purpose of improving the performance of the HLS module (*Edgington et al., 1996a*). Generally, the complexity of the text analysis problem depends on the depth of the orthography (*Bird, 1998*) of the language being analysed as well as the domain of the written document and the style of writing.

The information extracted during the text analysis is required for prosody modelling. The ease and accuracy with which such information can be extracted from text is therefore a crucial task. For example, the traditional Chinese and Thai orthographies do not use explicit sentence delimiters (*Shih and Sproat, 1996; Mittrapiyanuruk et al., 2000*). Hence, sentences have to be extracted from input texts during the segmentation process. The segmentation process determines the words and phrase boundaries as well. This information, particularly the phrase boundary, is used to determine the position of pause during the low level synthesis stage. This necessitates the need for a tokenisation process which, together with text segmentation process, extract a set of tokens with at least one space in between.

Other issues relate to how the TTS system handles delimiters when analysing the input text. In the IBM TTS system (*Sproat et al.*, 1996), for example, each sentence is viewed as consisting of one or more words separated by word delimiters. It is then expected that each word is separated from the other by a whitespace or other delimiters. Often a *Tab* indicates a new paragraph although *Klatt* (1987) has shown that this is not a reliable cue for paragraph. It is also important to note that the end of sentence maker, i.e. the period or full-stop, can also indicate abbreviation, a decimal point at the beginning or end of a number or a delimiter for name initials.

Similar difficulties are encountered when dealing with other delimiters such as commas (,), question marks (?), double quotes (“”) as well as apostrophe and single quotes (‘). It is important to note that SY uses similar method to delimit word boundaries as European languages. The task of text segmentation is much more complex in the Asian language TTS system. Like SY, the orthography of some Asian languages sometimes incorporate tone information and it is relatively easy to identify a syllable together with its associated tone.

Unlike European languages, Asian languages do not use spaces to delimit words. For example, in Mandarin Chinese, *Shih and Sproat* (1996) had to pre-process text by inserting spaces between words. The process introduce the need to use other symbols in addition to possible punctuation marks, as a delimiter of words. This results in the need to develop a more complex text segmentation algorithm than that obtained in TTS for European languages.

Cantonese orthography is similar to Mandarin. *Lo et al.* (1998) proposed a Cantonese text normalisation process in which special expressions, for example numerals, abbreviations, punctuation marks, are not expanded into their textual equivalent, as usually done with English and Mandarin. They are embedded in the tokenisation module and grapheme-to-phoneme (G2P) conversion process. The tokenisation module draws out the special symbol as a single token while the G2P conversion assigns a phonological representation according to the text features.

The text analysis problem in SY is much simpler than those described for Chinese, Thai and Cantonese above. In the case of SY, there are two subtasks that must be accomplished by the text analysis process. These are discussed as follows:

**Text segmentation.** In the text segmentation task, characters and groups of characters in an input text are assigned labels which identify them by type. An important task for the text segmentation module is the handling of the ambiguities associated with the use of punctuation marks. To illustrate this problem, consider the SY sentence “David fún mi ní ₦200.10k lánàá.” (meaning “[*David give me ₦200.10k yesterday*] *David gave me ₦200.10k yesterday.*”) The second full-stop in the sentence indicates the end of the sentence while the first one is a separator for the integer and fraction part of a currency. In a more complex sentence, a full-stop can be used to indicate a separator for letters in an abbreviation, such as N.C.B. (Nigerian Cocoa Board). Therefore, the segmentation module must be able to disambiguate punctuation and symbols used in similar manner by identifying and assigning them an appropriate label.

**Text normalisation.** The text normalisation task is perhaps the aspect of HLS module that presents the most challenging design problem. During the text normalisation process, all non-text items such as numbers, abbreviations, symbols, and acronyms must be expanded into their lexical (or textual) equivalent. This module is also responsible for converting textual anomalies such as foreign names, e.g. David, into a form amenable to pronunciation and with the correct accent in the target language. The problem in most African language text is that they contain a lot of foreign words and proper names from non-African languages. The need to determine a proper pronunciation for these foreign words imposes a lot of constraints on the design of an efficient text normalisation system.

Another text normalisation problem is currency expansion. Numbers representing currency values must be arranged and expanded in the format in which they are usually pronounced. For example, the currency specification “₦200.10k” in the above sentence must be expanded such that the value of the integer part, i.e. 200, must precede the currency symbol name, i.e. ₦, while the name of the fraction symbol name, i.e. k, must precede the fractional value, i.e. 10. The lexical expansion of the currency is therefore “Igba Náírà àti Kòbò Mèwàá” (meaning “[*Two hundred Naira and Kobo ten*] *Two hundred Naira and ten Kobo*”). The complexity here is that, this rule is not consistent and whether the cur-

rency units are pronounced before, in the middle, or at the end of the currency depends on the numerical value of the money. For example ₦5 is pronounced “Náírà Máraùún.” (meaning “[Naira five] Five Naira.”). In addition, the number must be expanded using the vigesimal (base 20), rather than the commonly used decimal (based 10) number system. Other text normalisation problems include: date expansion, special character processing, as well as acronyms and synonyms expansion.

The above discussion suggests that the text analysis problem in SY is not a trivial one. Often this problem can be solved by using a set of rules and/or tables which evaluates cues such as capitalisation of adjacent letters such as A.B.C. which may occur in text (Donovan, 1996), or adjacent numerals in the input sequence. For example, rule-based (Yao *et al.*, 1990) and dictionary-based (Liang and Zhen, 1991) approaches have been proposed and used. The problem with these approaches is that the effort required in constructing and maintaining a consistent rule-base that covers a reasonable scope is prohibitive.

Other solutions to the text analysis problem has been suggested for non-tone languages. This includes expert system approach for British English and French (Divay and Vitale, 1997), and *Finite State Transducer* (FST) for Greek (Yiourgalis and Kokkinakis, 1996), and Welsh (Williams, 1994) amongst others. Sproat *et al.* (1996) have also presented a text analysis system for TTS based on *Weighted Finite State Transducer* (WFST). Their approach has been applied to eight languages: Spanish, Italian, Romanian, French, German, Russian, Mandarin and Japanese. These approaches have also been applied to tone languages including Cantonese and Mandarin Chinese (Shih and Sproat, 1983, 1996) and Thai (Mittrapiyanuruk *et al.*, 2000; Tarsaku *et al.*, 2001).

The data-driven approach has also been put forward by a number of researchers in Chinese TTS (Breiman *et al.*, 1984; Dunning, 1993; Chiang *et al.*, 1996; Sproat *et al.*, 1996). Mittrapiyanuruk *et al.* (2000) presented an algorithm for extracting sentences from Thai text paragraphs by detecting the true sentence breaking spaces, using statistical part-of-speech (POS) analysis. Although this approach does not suffer from the problem of rule inconsistency, it is difficult to generate data that will have acceptable levels of coverage for the TTS task. In order to merge the strengths of the

rule-based and data-driven approaches, hybrid approach has been proposed (*Nie et al.*, 1995) which integrate the two techniques. The hybrid approach has achieved a higher accuracy and its research focus is on improving the performance.

Recently *Yu and Huang* (2003) defined the problem of text segmentation in Mandarin Chinese as a disambiguation problem and proposed a method based on a three-layer classifier (TLC). The first layer uses pattern table and decision tree. The next two layers are only visited if the first layer fails to resolve the ambiguity. The second layer uses Bayesian theory and adopts the voting scheme to compute the disambiguation score. Based on the algorithm's confidence of sense disambiguation, the third layer may exploit an alternative model to enhance the performance. Yu and Huang reported an accuracy of 99.8% and 97.5% for their training and test sets respectively.

The text analysis problem is beyond the scope of the present work. The text analysis problem discussed here highlights the task required in analysing SY text. The aim of this is to show that it is relatively easy to extract the information required for modelling prosody from SY text. In SY orthography, syllables and their associated tones are explicitly specified. This information can be marked up using modern text mark-up language such as XML. The task of extracting prosody information from mark-up text is less complicated when compared with plain text. We have developed a text annotated system in which the important information for SY prosody modelling are marked-up in the text using appropriate tags. A review of literature on text mark-up system is presented in Section 6.5. The detail development of our text mark-up system for SY text-to-speech application is in Appendix C.

### 2.2.3 Review of techniques in Low Level Synthesis (LLS)

The LLS process converts the output of the HLS into streams of sound which is expected to be intelligible and natural. Modern LLS synthesis methods can be grouped into three classes: articulatory (*Parthasarathy and Coker*, 1992; *Beskow*, 2004), formant (*Raptis and Carayannis*, 1997; *Jitca et al.*, 2002) and concatenative (*Wu and Chen*, 2001; *Lee and Cox*, 2002) speech synthesis methods. The conceptual framework underlying articulatory and formant speech synthesis method is the source-filter model. In this model, a signal excitation source drives a signal filtering mechanism which in turn

produces a speech waveform as output. This is a linear model without a feedback mechanism.

Concatenative speech synthesis methods, however, is developed around a functional model of speech production. The function expected to be performed by a speech synthesis system is replicated by modelling a system that does not necessarily imitate the original speech production mechanism. In the concatenative speech synthesis (CSS), the output of the HLS is used for selecting and joining together pre-recorded pieces of natural speech sound from a speech corpus. A brief review of literature on the three LLS synthesis methods is presented in the following subsections.

### Articulatory Speech Synthesis (ASS)

In articulatory speech synthesis (ASS) method, the aim is to derive a model (usually mathematical) of the human natural articulators namely: lips, tongue, oral, and nasal cavities, etc. The model is either a two- or three-dimensional model of the speech organs that are under the control of a set of deformation parameters (*Cohen and Masaro, 1993; Beskow, 1997; Birkholz and Jackèl, 2003*). The model is used to simulate the various configurations that the human speech organ can attain during speech production. The shape of the vocal tracts defined by the positions of the articulations is converted into transfer functions usually by estimating area function and formant frequencies (*Coker and Fujimura, 1966; Cohen, 1995*).

*Klatt (1987)* suggested that the vocal cord model might be similarly used to generate an appropriate excitation signal. Thus, the synthesis problem is reduced to specifying articulatory targets for each speech unit, e.g. phonemes, and accurately modelling the articulator dynamics. In a recent work, *Beskow (2004)* described an articulatory control model using ten parameters to produce control trajectories to govern articulatory movements for a given phonetic target specification. For example, a sequence of time-labelled phonemes, optionally including stress and phrasing markers. The results of the objective evaluation shows that a rule-based implementation of the model produced an intelligibility of 81.1%. Although the data-driven implementation of the model produced trajectories that best matches the targets, this advantage did not manifest in the intelligibility quality.

A major problem with the use of the ASS method arises during the modelling of speech prosody (*Wouters and Macon, 2002*). In intonation modelling, for example, there are many articulatory constraints on the production of  $f_0$  which are not directly reflected in the speech signal. For instance, in SY speech production, it is well known that certain consonants raise or lower the  $f_0$  of adjacent vowels (*Hombert, 1978*). The extent to which this constraint is accounted for by the articulators remains a major research problem in articulatory phonetics (*Xu and Sun, 2002*).

Also, the correlation between the mode of glottal excitation and the behaviour of the upper articulators, especially at abrupt segmental boundaries, are still issues of vigorous debate (*Xu and Wang, 2001; Xu and Sun, 2002*). As a result of the above stated problems, most articulatory speech synthesis systems use an oversimplified model of the physical properties of the human vocal tract mechanism. This has generally resulted in poor speech quality. The major reason is that an adequate model of the human vocal tract will result in a computationally complex and expensive system with low efficiency.

Many attempts have been made to solve these problems. For example, *Parthasarathy and Coker (1992)* presented an analysis by synthesis scheme for estimating the phoneme-level articulatory parameters to obtain best fits to natural speech, in the context of a text-to-speech (TTS) system. The working units of optimisation are the parameters of an articulatory model (one vector per phoneme) and vectors of time and speed of transition for each parameter.

ASS method has enjoyed little application and less attention in speech synthesis research due to the complexities involved in producing accurate articulatory model. The complexity arises from the inadequate theoretical framework for relating speech sound with the natural speech production mechanism. It has been argued by phoneticians that articulatory models cannot account for all the variability found in natural speech (*Kelly and Local, 1986; Cahn, 1998*). Recent effort at further increasing the accuracy of articulatory trajectories based on X-ray data using the Hidden Markov model (HMM) and Artificial Neural Networks (ANN) (*Blackburn and Young, 2001; Beskow, 2004*) has produced promising results.

The motivation for using ASS method in TTS is that articulatory parameters are likely to interpolate over long intervals, and reproduce many acoustic details from



simple constraints governed by physical laws. For text-to-speech applications, it may be easier to formulate rules in the articulatory domain. However, the state of the art in ASS does not allow for the development of a practical TTS, particularly for tone languages, where finer control on intonation contour generation is required.

### Formant Speech Synthesis (FSS)

In Formant Speech Synthesis (FSS), an attempt is made to bypass the complications of modelling articulatory movements and computing the acoustic effects of those movements as done in the ASS method. Therefore, the spectral properties of the transfer function are computed directly from linguistic representation using a set of carefully selected rules. The rules are generated in conjunction with a phoneme or syllable string specification of the desired utterance. Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech.

The first attempt at constructing a rule-based formant synthesis system was by *Kelly and Gerstman* (1961). Another successful effort is reported by *Holmes and Russell* (1999). These efforts produced very poor synthetic speech but they laid the foundation for modern FSS based systems. Research into formant timing and the perceptual relevance of specific spectrographic pattern (*Hertz and Huffman*, 1992) has led to improved models, for formulating rules that predict formant pattern. These models have not only resulted in simpler and more general rules for predicting formant trajectories in a particular language, but also in the expression of a wide range of language-universal patterns (*Kelly and Local*, 1986).

For example, *Lee et al.* (1993) presents a set of tones concatenation rule for a formant-based Mandarin Chinese TTS model (see Figure 2.5). Their concatenation rules are obtained empirically by carefully analysing the tone pattern behaviour under various tone concatenation conditions for many sentences in a database. Their approach used orthonormal expansion, vector quantisation, and statistical analysis. The system is Linear Predictive Coding (LPC) based, with all pre-recorded syllables represented and synthesised by LPC coefficients. They reported an intelligibility and naturalness score of 96.0% and 95.9% respectively.

The rules for formant synthesis are generally derived through an iterative process of

Illustration removed for copyright restrictions

Figure 2.5: Structure of the formant synthesiser used in *Lee et al.* (1993)

rule formulation in accordance with an underlying model. During the iteration process, the rules are refined through evaluation by auditory comparison with natural speech. The process has been automated by the use of special interactive rule development tools, such as the Delta System (*Hertz and Zsiga, 1995*). In an approach proposed by *Jitca et al.* (2002), fuzzy logic was used to model the rule base for the formant synthesis of Romanian. The approach predicts the characteristics of the first two formants, i.e.  $F1$  and  $F2$ , based on an analysis of phoneme behaviour.

Formant synthesised speech can be very intelligible, even at very high speeds, avoiding the acoustic glitches that can often plague other speech synthesis methods. Also, because formant-based systems have total control over all aspects of the output speech, a wide variety of prosody characteristics can be generated, conveying not just questions and statements, but a variety of emotions and tones of voice. That results in the high intelligibility of FSS based TTS systems. However, the naturalness quality of synthetic speech generated by FSS is not good enough as it often sounds robotic. Since our goal is to produce intelligible synthetic speech with maximum naturalness, the formant synthesis approach is not suitable for our purpose.

### Concatenative Speech Synthesis (CSS)

Concatenative Speech Synthesis (CSS) was introduced to try to bypass the problems encountered in trying to model the natural speaking process. In this method, a database of pre-recorded speech segments is created and the speech units are labelled. The

labelled units are concatenated together, in the time domain, to synthesise the desired utterance. Concatenative speech synthesis is the most popular approach in current speech synthesis research. This is because of its simplicity and the quality of its output, which is better than those of articulatory and formant speech synthesis methods. CSS also makes it possible to incorporate a wealth of information, particularly regarding co-articulation, which is very difficult to model using the other methods.

An ideal speech synthesis unit should facilitate the concatenation and production of all possible utterances without the need for signal processing. It is, however, impossible to record all speech units in all possible context. Usually, the prosodic features of the recorded units are different from the prosody in the target utterance. The units must therefore be selected from disjoint phonetic context. Hence discontinuities in spectral shape as well as phrase mismatches cannot be totally avoided (*Dutoit and Leich, 1994*). This results in the discontinuity of signal at the concatenation points leading to audible glitches in the synthetic speech. The art in CSS is to select units such that the glitches between adjacent units are minimised.

Theoretically, the larger the size of the speech database, the better the quality of synthetic speech. This theory, together with the availability of the technology for recording and processing large amount of data, resulted in the emergence of corpus-based approach in CSS (*Bellegarda et al., 2001; Chou et al., 2002; Iida, 2002*). In this approach, a large body of speech called *speech corpus*, produced by a single speaker, is collected. The corpus is designed with the aim to include almost all linguistic and prosodic features for the target language.

*Chou et al. (2002)* have argued that a parallel analysis of all prosodic features in the speech signal, together with corresponding texts, can lead to better prosody models. The results of the qualitative evaluation of implemented TTS systems, however, suggest that this approach is unlikely to produce the desired high quality prosody model. *Möbius (2003)* attributed this problem to the complexity and combinatorics of language and speech in general, and to the Large Number of Rare Events (LNRE) in particular. *O'Brien (1993)* stated that the problem is a result of natural speech being infinitely variable while databases are always finite.

At the signal level, a number of signal processing algorithms have been proposed

for prosody modification and smoothing acoustic units in CSS. They include: Linear Predictive Coding (LPC) and Pitch Synchronous Overlap (PSOLA) (*Chapentier and Stella, 1986*), Time-Domain PSOLA (TD-PSOLA) (*Moulines and Charpentier, 1990*), Multi-Band Resynthesis PSOLA (MBR-PSOLA) (*Dutoit and Leich, 1993*) methods as well as algorithms which employ harmonic models (*O'Brien and Monaghan, 2001; Takano et al., 2001*). Each of these methods have their strengthes and weaknesses and their application depends on the amount of distortion that can be tolerated in the final synthesised speech. For example, the TD-PSOLA method has been described by *Dutoit and Leich (1993)* as an intermediate between the two extreme situations, none of which offers satisfactory speech synthesis result.

More modern methods rely on automatic learning techniques employing stochastic tools (*Venditti and van Santen, 1998*), neural networks (*Sakurai et al., 2003*) or decision trees (*Donovan, 2003*). Such techniques automatically extract prosodic regularities from speech corpora, such as fundamental frequency (*Ross, 1995; Dusterhoff and Black, 1997*), segment duration (*van Santen, 1994; Shih and Ao, 1997*), or pauses (*Barbosa and Bailly, 1994*).

One of the disadvantages of the automatic learning approach is that it requires preliminary manual labelling of the speech corpus. This process is rather time consuming and tedious. There have been a few promising attempts to achieve automatic labelling. For example, *Campione et al. (1997)* proposed a stochastic technique which, when applied to prosody, seems to produce promising results. A conceptual limitation of automatic data-driven approaches is that it tends to mask out the linguistic relationships between the different levels of speech by replacing them with less accurate stochastic relationships. *Véronis et al. (1998)* observed that, at an optimal level, stochastic models reach a cutoff point above which performance no longer improves. Therefore, we conclude that corpus-based approach, though useful, has its problem.

Recently, *Wouters and Macon (2002)* proposed an approach for improving prosodic modification of acoustic units, by controlling the degree of articulation of sonorant phonemes. The modifications were based on numerical data obtained from the effects of prosodic factors on the spectral dynamics for balanced corpus. The modification technique, discussed in Wouters and Macon's work, showed that it is possible to alter

the formant structure of acoustic units while preserving the perceptual quality of the original recorded speech. The technique thus increases the flexibility of concatenative speech synthesis systems, and opens up new possibilities to explore the perceptual effect of quantitative formant modification in natural-sounding speech.

Despite the sophisticated techniques employed in the design of these algorithms, the problem of low quality prosody still plagues CSS. It seems, therefore, that an approach employing an holistic view with a strong theoretical motivation that can be implemented using computational mechanism is required.

## 2.3 Unit of speech synthesis

The speech synthesis approach adopted in this work is *concatenative speech synthesis* (CSS). It involves taking pre-recorded speech signals and *concatenating* (i.e. joining head-to-tail) them. In the CSS approach, the size and type of *synthesis unit* that forms the element of concatenation is a very important design issue (*Hunt and Black, 1996*). This is because the *synthesis unit* constitutes the basic building block for utterances and it affects the overall quality of the resulting synthetic speech output. When designing or selecting a synthesis unit, it is important to take the linguistic and orthographic features of the target language into account.

Speech synthesis units can generally be classified into two broad groups namely: (i) lexical and (ii) sub-lexical<sup>1</sup>. The lexical speech synthesis units are composed of words or phrases. Units longer than words are not used in modern Text-to-speech (TTS) system. This is due to their inflexibility as well as the difficulties involved in manipulating the parameter of longer units in synthesis of diverse utterance. Also, the large number of such units make their collection and annotation very complex, even for small and restricted TTS application. Therefore, the use of units longer than a word is impractical.

The word has been suggested and used as the basic unit of speech synthesis (*Olive and Nakatani, 1974; Eric and Tatham, 1999*). The advantage of using word is that the rules for synthesising speech by word concatenation are relatively simpler since no rule is required to account for the intricate intra-word co-articulation and timing

---

<sup>1</sup>i.e. sub-lexical are units that are shorter than words.

Table 2.1: Synthesis Units in some popular TTS

Systems	Synthesis Unit	Main language(s)
Festival ( <i>Black et al.</i> , 1999)	Diphone/non-uniform units	British English
Donovan ( <i>Donovan</i> , 2003)	Diphone	British English
MBROLA ( <i>Dutoit and Leich</i> , 1993)	Diphone	British English
<i>Lee et al.</i> (1983)	phone	Mandarin
<i>Shih and Sproat</i> (1983)	triphone	Mandarin
<i>El-Imama</i> (1990)	Demi-syllable	Arabic
<i>Lee et al.</i> (1993)	syllable	Mandarin
<i>Wu and Chen</i> (2001)	syllable	Mandarin
<i>Kochanski et al.</i> (2003b)	syllable	Mandarin
<i>Chan and Chan</i> (1992)	syllable	Mandarin
<i>Chou et al.</i> (2002)	syllable	Mandarin
<i>Mittrapiyanuruk et al.</i> (2000)	syllable	Thai
<i>Wang and Hwang</i> (1993)	syllable	Taiwanese
<i>Lee and Oh</i> (1999)	syllable	Korean
<i>Lo et al.</i> (1998); <i>Lee et al.</i> (2002a)	syllable	Cantonese

effect. However, storing all possible words in the context within which they can occur is highly prohibitive.

This forces researchers to consider smaller units. Such units include: phones (*Lee et al.*, 1983), diphones (*Isard and Miller*, 1986; *Donovan*, 1996; *Hunt and Black*, 1996), triphones (*Shih and Sproat*, 1983) and demi-syllable (half-syllable) (*El-Imama*, 1990) (see Table 2.1).

The *phone* has been a very attractive unit for speech synthesis due to their very small number, for example, there are about 44 phonemes in British English (*Edgington et al.*, 1996b). The major problem with phones is that storing one example phone for each phoneme in a language will not produce good quality synthetic speech. This is because speech phenomena, such as co-articulation, cause the production of phones to vary due to the characteristics of neighbouring phones. Since the number of phones in a language can be as much as 50, for example English has 44, storing all possible phones for all possible contexts is a very complicated task which will also require a large inventory of synthesis units. Even then, the resulting database will not guarantee accurate speech units for producing smooth concatenation.

The *diphone* represents a computational expedient compromise between the most

immediate effects of co-articulation and the desire to reduce the number of unit stored. The use of diphone as synthesis unit pre-supposes two fundamental assumptions. The first assumption is that phones can be decomposed into three components, namely: onset, steady-state and offset. The second assumption is that the effects of co-articulation can be captured within the transition from one phone to another. It has been shown that these assumptions are not accurate when applied to tone language speech. For example, in Mandarin Chinese synthesis, *Shih* (1995) discovered that the formant values of glides, which have short fast-moving formant trajectories, are heavily influenced by the following vowel when diphone based synthesis units are used. A similar problem was also experienced in the synthesis of central vowels such as /e/.

A better approximation than the diphone can be obtained by defining a unit that contains the transition from one steady-state portion of the neighbouring phones into and out of a phone. Such a unit is the *triphone*. However, in order to obtain an acceptable synthetic speech quality, the inventory of triphone can be prohibitively large. For example, in speech synthesis for a southern British English dialect, the triphone model results in a total of 38,892 units (*Edgington et al.*, 1996b). This is quite a large number when compared with the 44 phoneme units for the same language.

In order to address the problems associated with the use of the speech synthesis units discussed above, some researchers have suggested the use of half-syllable or *demi-syllable* (*El-Imama*, 1990). The immediate problem that arises with the demi-syllable is the difficulty in defining the constituent of a whole syllable and the partitioning of the acoustic signal of a whole syllable into two equal halves. The limitation of the above sub-lexical unit may be as a result of their weak linguistic status. Apparently, the diphone, triphone and demi-syllable are not linguistically motivated units since no account of their underlying structure of language has been convincingly constructed. Another approach suggested in the literature is the combination of speech units, for example, phone and demi-syllable (*Kiat-arpakul et al.*, 1995) and multiform units (*Takano et al.*, 2001). However, this approach introduces more complications into the speech unit selection process and makes the unit selection algorithm less effective.

The *syllable* has been suggested and used as unit in speech synthesis, particularly for tone languages. The notion of a syllable is self-evident to native speakers of a tone

language. Take the SY sentence “Adé àti ìgè ló kókó ra ìwé Atóka” (meaning “[Ade and Ige the first buy book Atoka] Ade and Ige are the first to buy the book Atoka”) for example. When spoken at a normal speech rate, the syllables in the sentence, which are fifteen in number, as well as the tones associated with each syllable are distinctly perceptible. This type of “definitive” specification of a syllable still remains a complex and controversial issue when dealing with non-tone languages such as English (Kessler and Treiman, 1997; Alamolhoda, 2000). This probably accounts for why TTS researchers in many non-tone languages do not consider the syllable in the same manner as those working on tone languages.

There are two factors which influence our selecting of the syllable as the *basic unit* of speech synthesis. One of them is the fact that the syllable is used in most of the tone languages we reviewed in the literature (e.g. Wu and Chen (2001); Chen et al. (2003)). The second is a critical analysis of literature in the phonetics and phonology of SY coupled with the experience of a native speaker of the language.

Our selection of the syllable as the basic unit for SY synthesis is grounded in the understanding that the syllable is the fundamental entity in the production and perception of SY speech. The syllable is also properly wedded to the higher tiers of linguistic organisation of SY utterance. Furthermore, the syllable is a perceptually and acoustically coherent unit (Adéwolé, 1988) and that the syllable can be considered as the basic unit of prosody in tone languages.

Our choice of the syllable as the unit for speech synthesis is consistent with other work on TTS for tone languages such as Mandarin Chinese (Lee et al., 1983; Chan and Chan, 1992; Lee et al., 1993; Wu and Chen, 2001; Kochanski et al., 2003b), Thai (Mittrapiyanuruk et al., 2000), Taiwanese (Wang and Hwang, 1993) and Cantonese (Lo et al., 1998; Li et al., 2004) (see Table 2.1). This suggests that using the syllable as the basic unit in speech synthesis and processing, at least for tone languages, is not only intuitive but also a practical choice.



## 2.4 Summary

In this chapter, we have presented a brief history and discussed the state-of-the-art in modern speech synthesis. We reviewed various methods for speech synthesis including articulatory, formant, and concatenative speech synthesis methods. We selected the concatenative speech synthesis method as the basis for our prosody modelling because of its simplicity and the reported high quality of the synthetic speech produced by the method. We also reviewed and discussed the various units for speech synthesis including word, phone, di-phone and syllable. We selected the syllable as our unit of speech synthesis because the syllable is a perceptually and acoustically coherent unit.

# Chapter 3

## The standard Yorùbá language

### 3.1 Phonology

Numerous linguistic works (*Maddieson*, 1984; *Lindblom and Engstrand*, 1989; *Pellegino and Andre-Obrencht*, 2000) have shown that languages can be characterised, to a large extent, by their phonological systems. The fundamental element of the phonological system is the inventory of phones pronounced by native speakers of the language. For tone languages, this system can be split into a *vowel system*, a *consonant system* and a *tone system*. The inventory of the sounds present in spoken utterances is a basic process in the definition of the phonological structure of the speech production for a language. The Standard Yorùbá consonant and vowel systems are shown in Table 3.1 and 3.2, respectively (*Ògúnbòwálé*, 1970; *Owólabí*, 1998).

The SY alphabet has 25 letters which is made up of 18 consonants (represented by the graphemes: *b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ʃ, t, w, y*) and seven vowels represented by the graphemes: (*a, e, ɛ, i, o, ɔ, u*) (*Adéwolé*, 1988). Note that the consonant *gb* is a diagraph, i.e. a consonant written in two letters. There are five nasalised vowels in the language (*/an/, /en/, /in/, /ɔn/, /un/*) and two pure syllabic nasals (*/m/, /n/*).

SY has three phonologically contrastive tones: High (H), Mid (M) and Low (L). Phonetically, however, there are two additional allotones or tone variant namely, rising (R) and falling (F) (*Connell and Ladd*, 1990). A rising tone occurs when an L tone is followed by an H tone, while a falling tone occurs when an H tone is followed by an L

Table 3.1: Segmental phonemes of Standard Yorùbá consonant

Manner of articulation	Place of articulation						
	Bilabial	Alveolar	Palato-Alveolar	Palatal	Velar	Labio-Velar	glottal
Stops	b	t d			k g	kp gb	
Fricate	f	s	ʃ				h
Affricates				j			
Nasal	m	n					
Flap		r					
Lateral		l					
Semi-vowel	w			j			

tone. This situation normally occurs during assimilation, elision or deletion of phonological object as a result of co-articulation phenomenon in fluent speech (*Bámgbóṣé*, 1966; *Akinlabí*, 1993; *Déchaine*, 2001).

Table 3.2: Segmental phonemes of Standard Yorùbá vowel

Oral vowels	Nasal vowels

### 3.2 Phonological structure of Yorùbá syllable

Linguists have proposed several theories for possible configuration of the internal structure of syllables (*Kessler and Treiman*, 1997). The fundamental differences amongst these theories are in the terminology of basic components and how those components are organised in the syllable. In the *moric* theory of syllable structure, for example, the

components of the syllable are units of weight called moras (*Hayes, 1989*). The basic moric theory always has the vowel as the first mora and the coda as the second, with the complication that a long vowel is considered to be simultaneously in both moras. In the *body-coda* configuration theory, the vowel is grouped with the onset to form a constituent called the body (*Inverson and Wheeler, 1989*).

In the *flat syllable* theory (*Clements and Keyser, 1983*), a Consonant-Vowel-Consonant (CVC) syllable is made up of a vowel, an onset and a coda. The vowel is isolated from both the onset and the coda. In the *onset-rhyme* theory, on the other hand, the vowel is grouped with the coda to form a constituent called the rhyme (*Goldsmith, 1990*).

After a careful review of the literature on SY phonology and phonetic (*Bámgbóṣé, 1966; Adéwólé, 1986*), we observed that most, if not all, phonotactic constraints in SY involves the vowel. The vowel can be preceded by a consonant, which is the onset, and followed by the nasal /n/ in the coda position. We therefore analyse SY syllables using the *ONSET-RHYME* theory (*Goldsmith, 1990; Greenberg, 1998*). The “ONSET” occupies the first position in the syllable structure. If present, it is any of the SY consonant (i.e. *b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ʂ, t, w, y*). The “RHYME” occupies the rest of the syllable and it is either a vowel, a nasalised vowel or a syllabic nasal.

The RHYME is further divided into *NUCLEUS* and *CODA*. The nucleus is the central element of the syllable and all syllables are built around a nucleus. In phonetics and phonological theory of language, there are two types of nuclei: *vocalic* and *non-vocalic* (*Kessler and Treiman, 1997; Manteescu, 2004*). A vocalic nuclei in Yorùbá can be any of the seven vowels (*a, e, ɛ, i, o, ɔ, u*).

The vowel in nasalised vowels forms vocalic nucleus. Non-vocalic nuclei are associated with syllabic-nasal and they can be any of the two nasal consonants /n/ and /m/ or their allophones. The coda, when present, is the consonant /n/. The ONSET is optional, that is some syllables do not have an ONSET. However, all syllables have a RHYME with which its tone is associated. The CODA is not present in some syllables made up of an isolated vowel, a combination of a consonant and a vowel, or syllabic nasal. Since a syllabic nasal has a non-vocalic nucleus, both the ONSET and CODA are not present.

Based on the above analysis, we designate five types of syllable configurations in SY (see Figure 3.1). This includes: Consonant-Vowel (CV), Vowel (V), Consonant-Nasalised Vowel (CVn), Nasalised Vowel (Vn) and Syllabic Nasal (N). Many linguistic phenomena in SY, particularly the prosodic attributes of utterances, are easily described in terms of the properties of the RHYME.

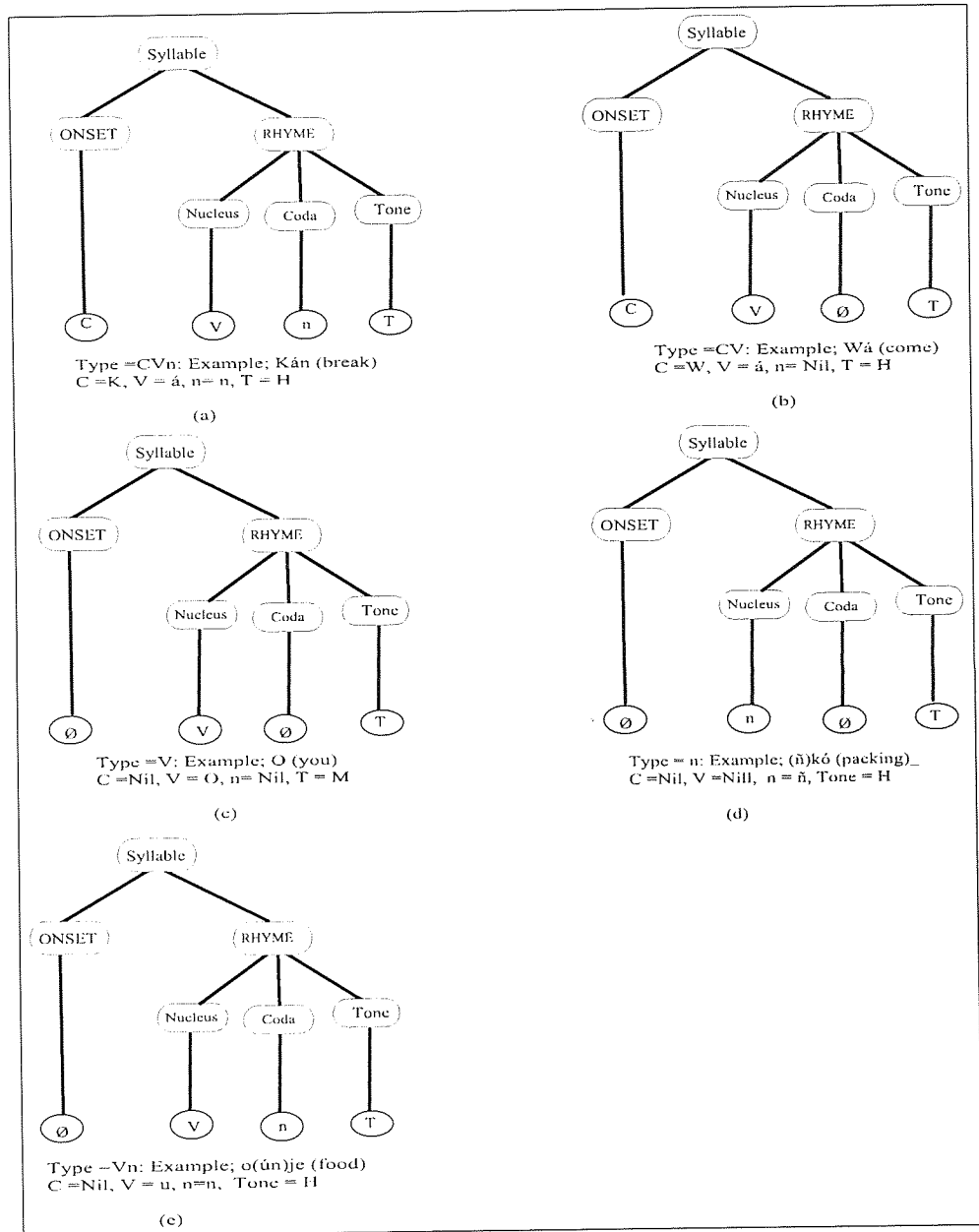


Figure 3.1: SY syllable structures with examples (Note:  $\emptyset$  implies Nil)

### 3.2.1 Syllable inventory

Using the above syllable configurations, we have generated a syllable inventory for SY speech synthesis using the following procedure. Since there are 18 consonants in SY, there can be a total of 18 ONSETs of consonant type. All possible combinations of consonant and vowel as well as consonant and nasalised vowels form a valid syllable in SY.

Five of the vowels, namely: (/a/, /e/, /i/, /o/, /u/) can combine with the CODA consonant /n/ to form five nasalised vowels. If each of the consonants are combined with a vowel, we will have a total of  $18 \times 7 = 126$  CV type syllables. If each of the consonants are combined with a nasalised-vowel, we will have a total of  $18 \times 5 = 90$  CVn type syllables. SY also has two syllabic nasals /n/ and /m/.

This gives a total of  $126 + 90 + 14 = 230$  based syllables (see Table 3.3). The three lexical tones (cf. Section 3.1) can occur freely on syllables irrespective of their syllabic structure. Hence, we have a total of  $230 \times 3 = 690$  phonological syllables in SY. The inventory of the syllables based on their phonological structure is shown in Table 3.3. Table 3.4 shows the distribution of the phonological structures of SY syllables.

Table 3.3: Phonological structure of SY Syllable

Tone syllables (690)				
Base syllables (230)				Tones(3)
[ONSET] (18)	RYHME(14)			H, M, L
Consonant	Nucleus		Coda	
	Vocalic	Non-Vocalic	n(1)	
C	V (7)	N(2)		

It should be noted that although the CVn syllable configuration ends with a consonant, the consonant and its preceding vowels are the orthographic equivalent of a nasalised vowel. There is no consonant cluster in SY language and there is no closed syllable. The complete listing of all the syllables is provided in Appendix A.

Table 3.4: Statistics of Yorùbá syllables

Syllable Type	Syllable structure	Count		Percentage of total
		Base	Tone-syllable	
Vowel	V	7	21	3.04
Nasalised Vowel	Vn	5	15	2.17
Consonant Vowel	CV	126	378	54.79
Syllabic Nasal	N	2	6	0.87
Consonant Nasalised Vowel	CVn	90	270	39.13
Total		230	690	100.00

### 3.3 SY tone phonology and phonetics

In the surface *tonology* of Yorùbá, there are three contrasting tone phoneme or tonemes: (i) High (H), (ii) Mid (M) and (iii) Low (L). Each syllable has one tone associated with it. Phonetically, H is realised as Rise if it appears after L, and L is realised as Fall if it appears after H. These effects can be chained, so that HLH is realised as H-Fall-Rise. There is a widespread phenomenon of elision and contraction which sometimes results in the association of two tones with a vowel. This phenomenon can also create floating tones that are deleted post lexically (*Bámgbóṣé*, 1966; *Connell and Ladd*, 1990; *Akinlabí*, 1993).

As we move from the phonological to the phonetic aspect of speech, however, it will become increasingly difficult to model SY syllables in isolation from its associated tone. This is because an SY syllable cannot be pronounced in isolation of its tone. In order to allow for proper analysis during modelling, we view the syllable as an object (see Figure 3.2). The syllable object contains two main classes: ONSET and RHYME. The RHYME can further be decomposed into three subclasses: nucleus, tone and coda. The nucleus and tone subclasses are embedded and cannot be separated from each other. The ONSET class, nucleus and coda subclasses are phonetic entities and are sometimes referred to as the base syllable (*Wang and Chen*, 1994). The tone subclass, on the other hand, is a suprasegmental entity identified by its toneme. We refer to it

as the tone tier of the syllable object.

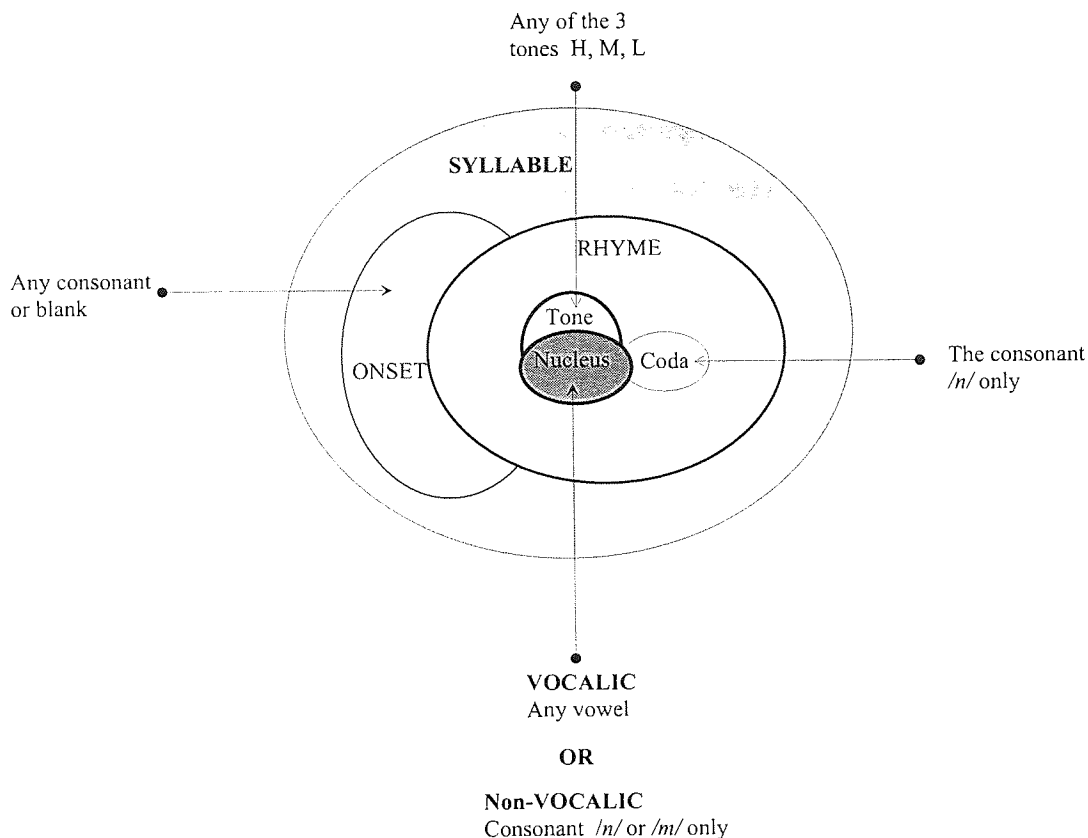


Figure 3.2: Yorùbá syllable viewed as an object

### 3.3.1 SY syllable and prosody

Most of the important prosody phenomena in SY can be explained using the syllables as the basic unit. The supra-segmental features of the syllable, which accounts for its prosody behaviour, is mainly determined by the syllable RHYME. However, some *phonotactic* and *tonotactic* constraints on syllable structure involve the syllable ONSET.

The acoustic correlate of the base tier are the first two formant frequencies (F1, F2). All the vowels can be easily identified by the F1 and F2 patterns in their speech spectrograms. The acoustic correlates of the tone tier is the fundamental frequency ( $f_0$ ) curve over the speech signal corresponding to the syllable. In SY, the  $f_0$  curve of a syllable spoken in isolation generally corresponds well with the canonical pattern of its tone (*Harrison*, 2000). In fluent speech, however, tones undergo both phonological and phonetic modifications due to *tone-sandhi* and *tone co-articulation*. These pheno-



mena cause the  $f_0$  curves of tones in fluent speech to deviate from the canonical form (Akinlabí, 1993).

Within the context of this thesis, therefore, the SY phonetic syllable is regarded as a segment of speech that is identified by continuous parameters organised around one local  $f_0$  peak and valley, and possibly preceded and/or followed by static segments. Whenever a syllable contains unvoiced parts, the tonal segment associated with it is taken to coincide with the voiced part of the syllable. This is to guarantee that the concatenation of all the tonal segments preserves the original  $f_0$  curve and Voiced/Unvoiced parameters. The  $f_0$  curve over syllables is an important element in the generation of the prosody for an utterance. An adequate speech database that is meant for modelling intonation of SY utterances should, therefore, include an accurate inventory of syllables and their  $f_0$  curve (preferably) described parametrically.

### 3.4 The standard Yorùbá orthography

An orthographic system for Standard Yorùbá (SY), using the Roman alphabet, has been in use since about two centuries (Bámgbóşé, 1965). A few extra diacritics are used to indicate additional sounds; for example, acute (´) and grave (`) accents indicate high and low tones, which are not present in Western languages. Orthographically, [ɛ] and [ɔ] are represented as *e* and *o* respectively. The labial-velar stops [kp] and [gb] are orthographically represented as *p* and *gb* respectively. The three fricatives are represented as *f*, *s*, and *h*. The palato-alveolar fricative [ʃ] is represented by the dotted consonant *ṣ*. The remaining consonants are sonorants: *m*, *l*, *r*, *y*, and *w* are written as shown in Table 3.1. The syllabic nasals are orthographically represented as *m* and *n* but their pronunciation is context dependent. If the following segment is a vowel, then the syllabic nasal is pronounced as a velar, e.g. “nògbò<sup>1</sup>” (meaning “[Not I hear] I didn’t hear”). When the syllabic nasal is followed by a consonant, the nasal is homorganic to the following segment, e.g. “mbò” (meaning “coming”), “ńso” (meaning “saying”). The syllabic nasal is generally written as ‘*m*’ when it precedes *b*.

Orthographically, the nasalised vowels are represented using a vowel+n sequence (i.e. *an*, *en*, *in*, *on*, *un*) when immediately following an oral consonant (e.g. “sin”

<sup>1</sup>This is a trisyllabic word made of the syllables *n*, *o* and *gbò*

(meaning “bury”) and “yàn” (meaning “select”). They are represented as a simple vowel when immediately following a nasalised consonant (e.g. “mò” (meaning “know”) and “nọ” (meaning “spend”). The other vowels are represented as they appear in Table 3.2. Apart from the diagraph /gb/ and the use of diacritic to mark tones on syllabic nucleus and dots used with some letters such as ẹ, SY written texts are similar to English texts.

The mid tone in SY is the default tone and it is represented orthographically with the dash or bar, -. It is not usually marked in most text. The two contrastive tones, i.e. Low and High are represented orthographically as grave (̀) and acute (́) respectively. The tones are associated with the RHYME part of a syllable and their diacritic marks are placed on the nucleus element bearing the tone. For example, in the word “Ikán” (meaning “ant”) (see Table 3.5) the first syllable, i.e. *I*, is a V type syllable carrying a mid tone. The second syllable is a CVn type syllable. The ONSET is the consonant *K* and the RHYME is *án*. The CODA is *n* while the nucleus vowel *a* carries the high tone diacritic mark, i.e. ́.

Table 3.5: Example of syllable structure for the SY word *Ikán*

WORD = Ikán							
Syllable = I			Syllable = kán				
ONSET	RHYME		TONE	ONSET	RHYME		TONE
	Nucleus	Coda			Nucleus	Coda	
	I		M	K	á	n	H

In Yorùbá tonology, HL sequence are realised as H1H2L. This is a phonetic representation, describing the raising of H before L. This phenomenon accounts for the occasional appearance of a sequence of two level tones on one syllable, creating rising or falling sequence tone. There are basically two situations in which the phenomenon can occur in the orthography. They are referred to and transcribed as: Low-High (LH) and High Low (HL). For example, the syllable *a* with LH sequence is written as *ǎ* and that with HL tone contour is written as *â*. We will however adopt the orthography in which each level tone is carried by exactly one syllable. Thus *ǎ* will be written as *àá* and *â* is written as *áà*.

All the punctuation used in normal English text also applies in SY text. Apart

Table 3.6: Punctuation in SY text

Type of Punctuation	orthography
Paragraph separation	New line, Multiple new lines
End of clause	Comma (,), semi-colon (;) and colon (:)
Quotations	double quote (“ ”), single quote (‘ ’)
Exclamation	exclamation mark (!)
Question	Question mark (?)
End of sentence	full-stop (.) and capitalisation

Table 3.7: Other tokens in SY text

Token type	Description/ general format
Date	String of number formatted as 99-99-99, 99/99/99, or a combination of number and letters
Time	String of number formatted as 99:99, 99/99
Currency	String of number prefixed by currency symbol, e.g. \$
Ordinal digits	String of number prefixed by a noun
Cardinal digit	String of number postfixes by a noun
Loanwords and proper names	Words and names written in English, French, Arabic, etc., spelling
Acronym	Group of upper case letters such as FIFA, OAU, USA, with or without separating full-stop marks.
Special Character	They are characters outside the above set, e.g., *, +, etc. usually intended for arithmetic expression

from punctuation marks, other tokens that can appear in an SY text include:

- foreign or loan words and proper names;
- numbers which represent currency, ordinal and cardinal digit;
- acronym, and
- special characters (see Table 3.7).

### 3.4.1 SY texts

An ideal SY text suitable for processing in the context of TTS system will be written using the orthography discussed in Section 3.4. In reality, this is not always the case.

There is a wide spectrum of writing styles and methods. At one end of the spectrum, the text is fully tone marked and all the under-dots are indicated. At the other end, the text is not tone marked at all and the under-dots are not marked as well. Between these extremes are a variety of text that uses the tone marks and under-dots at various degree of completeness. The contents of the written text, therefore, depend on the authors' ability to use the language as well as familiarity with the orthography. It also depend on the orientation of target readers as well as the purpose of the writing.

In general, most printed materials in Yorùbá language do not include diacritic marks. A reason for this, suggested by *Connell and Ladd* (1990), is that fully tone-marked Yorùbá text tends to put literate Yorùbá readers off and make reading more difficult hence most text are preferable written without the tone marks. Another important reason why most Yorùbá writers ignore tone marking is because it tends to slow down the speed of writing and hence obscure thought during writing.

With the advent of the computer technology, another reason is the inability of most users to use popular word-processor, such as *Microsoft Word*, for generating the required tone marks and under-dots for letters. Although Yorùbá letters with tone marks are not available on standard keyboard, some software, for example *Word Perfect*, can be manipulated to indicate tone marks on some vowels but most users do not know how to use them. The  $\text{\LaTeX}$  software is an ideal system for typesetting SY text since it contains all the features required for generating the tone marks and under-dots. However,  $\text{\LaTeX}$  is too complicated for an average user to learn and use within a short time.

A survey of four Yorùbá daily newspapers and four standard textbooks for teaching Yorùbá was conducted (see Figure 4.7 for examples of such text). The result, summarised in Table 3.8, shows that none of the local newspapers make use of tone marks in their writing whereas tone markings are used in Yorùbá language education textbooks in varying degree.

Due to the ambiguity involved in the analysis of partially marked and non-marked SY text, we assume and use fully marked SY text in the present work. Since all the non-default tones are marked and all under-dots are indicated, it is easy to assign tone to each syllable in the text. This makes the computation of pronunciation less complex.

Lotiito mo pada si ile awon Adeyemi, Tunde ko si fi mi sile ni igba kankan. gbogbo ibi ti e ba ti ri Tunde ni e o ti ri emi Debora, ni osu kesau-an odun ti o koja ni mo pe omo odun metafelogbon, ki n to de lati ibi ise ni ojo yii, won ti seto ojo ibi mi sile, n ko tile ranti mo, bi mo se wo ile ni mo ri gbogbo eniyan ti won n se ti won n so, ki fo n sele nile Adeyemi lonin, gbogbo won si pariwo *happy birthday*, Tunde sare so mo mi, o si yo oruka o fi si mi lowo, o ni oni ni ojo ibi re, ni oni yii gan -an ni

16 Olorun si da imole nla mejì; imole ti o tobi lati se akoso osan, ati imole ti o kere lati se akoso oru: o si da awon irawo pelu.  
 17 Olorun si so won lojo li ofurufu orun, lati ma tan imole sori ile,  
 18 Ati lati se akoso osan ati akoso oru, ati lati pala imole on okunkun: Olorun si ri pe o dara.  
 19 Ati asale ati owuro o di ojo kerin.  
 20 Olorun si wipe, Ki omi ki o kun fun opolopo eda alaye ti nrako, ati ki ciye ki o ma fo loko ile li oju-ofurufu orun.  
 21 Olorun si da erinmi nlanla ati eda alaye gbogbo ti nrako, ti omi kun fun li opolopo ni iru won, ati ciye abiyi ni iru re: Olorun si ri pe o dara.  
 22 Olorun si sure fun won pe, E ma bi si i, e si ma re, ki e kun iau omi li okun, ki ciye ki o si ma re ni ile.  
 23 Ati asale ati owuro o di ojo karun.  
 24 Olorun si wipe, Ki ile ki o mu eda alaye ni iru re jade wa, eran-osin, ati ohun ti nrako, ati cranke ile ni iru re: o si ri be.  
 25 Olorun si da cranke ile ni iru tire, ati eran-osin ni iru tire, ati ohun gbogbo ti nrako lori ile ni iru tire: Olorun si ri pe o dara.  
 26 Olorun si wipe, Je ki a da enia li aworan wa, gge bi iri wa: ki awon ki o si joba lori eja okun, ati lori ciye oju-orun, ati lori cranke, ati lori gbogbo ile, ati lori ohun gbogbo ti nrako lori ile.  
 27 Be li Olorun da enia li aworan re, li aworan Olorun li o da a; ati ako ati abo li o da won.

(a) SY text with no tone mark and (b) Partially tone marked SY text include foreign spelt words

**9 ÀWON IRÓ ÀTI ÈDÀ WỌN**

Ọ̀nà mejì pàtàkì nì a lẹ̀ gbà kọ̀ ẹ̀kọ̀ nípá àwọn iró èdè.

Ọ̀nà kúnf, a lẹ̀ kọ̀ ẹ̀kọ̀ nípá oríṣáṣí tábí onírúurú iró tí àwá èniyàn í pè jáde (ní ọ̀nà) nígba tí a bá n fẹ̀ lyeṣ nì pé kí a kọ̀ nípá bí a se máa nì gbé àwọn iró ẹ̀kọ̀ jáde, kí a se apẹjùwe àwọn iró bẹ́ẹ̀, kí a pín wọ́n sí ọ̀wọ̀ṣe abí. Nínú ẹ̀ka ẹ̀kọ̀ ẹ̀dà-èdè tí a nì pé nì fonqíki nì a tí máa nì kọ̀ nípá àwọn ohun tí a ká sílẹ̀ wọ́yí.

Ọ̀nà keji ẹ̀wé, a lẹ̀ kọ̀ ẹ̀kọ̀ nípá iró tí a nì lè nínú èdè kan láti fi iyáti hàn láàárín itumọ̀ ọ̀rọ̀ kan àti òmíràn, àwọn iró tí kò lè fi iyáti bẹ́ẹ̀ hàn ibátan tí ó wá láàárín àwọn iró tí ó lè fi iyáti hàn àti àwọn tí kò lè fi iyáti hàn láàárín itumọ̀ àwọn ọ̀rọ̀ àti báṣánlì tí a kò àwọn iró sí nínú èdè tí a nì lẹ́kàn. Àpapọ̀ ohun tí a ká sílẹ̀ yìí nì a mọ̀ sí ètò iró. Nínú ẹ̀ka imọ̀ ẹ̀dà-èdè tí a nì pé nì fonqíki nì a tí máa nì kọ̀ nípá ètò iró èdè kọ̀kkan gẹ́gẹ́ bí a yì sáláyé yìí. Bẹ́ẹ̀ sí nì fonqíki èdè kan a máa yáti sí tí èdè mílíràn.

Àmọ́ ọ̀dà, kí a tò lè se pé ẹ̀kọ̀ nípá iró èdè kún ojú iwọ́n tó, a gbéde lí sáláyé nì ọ̀rínkinníwó tró ibátan tí ó wá láàárín àwọn ọ̀nà mejìdẹ́jì tí a se p a lẹ̀ gbà kọ̀ ẹ̀kọ̀ nípá iró inú èdè yìí.

Nì tí èdè Yorùbá, a tí sáláyé ikínf nínú àwọn ọ̀nà mejìdẹ́jì tí a se pé a lẹ̀ gbà kọ̀ ẹ̀kọ̀ nípá iró èdè nì apá kúnf iwé yìí. Nì apá keji ẹ̀wé, a ó sáláyé ọ̀n keji náà nì kíkún. Síe iró àláyé bẹ́ẹ̀ yé ó sí mọ̀ kí a lè rí í bí ọ̀nà mejìdẹ́jì se ara kọ̀ ara pèlú.

9.1 Fòónù gẹ́gẹ́ bí iró asẹ̀yàti tábí iró aláṣeyàti

Nínú fonqíki tábí fonqíki, orúkọ̀ tí a máa nì se iró kọ̀kkan tí àw èniyàn nì pé jáde nígba tí a bá n fẹ̀ (lyeṣ, iró ife) nì fòónù. Ọ̀rọ̀lọ̀pọ̀ à onírúurú fòónù bẹ́ẹ̀ nì ó sí máa nì se yọ̀ nínú èdè kọ̀kkan. Àmọ́ ọ̀dà, kí a se gbogbo àwọn fòónù wọ́yí nì a lè lè láti fi iyáti hàn láàárín itumọ̀ ọ̀rọ̀ ka àti òmíràn nínú èdè kọ̀kkan. lyeṣ nì pé nínú èdè kọ̀kkan, àwọn fòónù ka

(c) Fully tone marked SY text (Owólabí (1998)).

Figure 3.3: Sample scanned texts in SY orthography

Table 3.8: A survey of text in Yorùbá textbooks and newspapers

Text type	Sample	Number of syllable counted	Number appropriate tone marked and under-dotted	% accurate orthography
Newspaper	Aláròyè	5,000	0	0.00
	Akéde	5,000	0	0.00
	Alálàyé	5,000	0	0.00
	Ìròyìn Yorùbá	5,000	0	0.00
Text Books	Ìjìnlẹ̀ Itúpalẹ̀ Èdè Yorùbá (1)	5,000	4900	98.00
	The Essentials of Yorùbá Language	5,000	4800	96.00
	Fonólóji àti Gírámà Yorùbá	5,000	4900	98.00
	Àsà Ìbílẹ̀ Yorùbá	5,000	3900	78.00

In the future, we hope to develop a technique that can automatically predict the tone pattern for any given syllable within a Yorùbá text. This technique, when developed, would be suitable for use as an interface to any SY text for the TTS system proposed here.

## Chapter 4

# A review of intonation models

The study of intonation has attracted a great deal of attention in speech and language science and engineering research. This attention arises from the conviction that intonation has an important communicative value. It has been argued (*Collier, 1990*) that intonation and other prosody feature of speech add something to the content and structure of speech sound which is not already expressed in the semantics of its individual words, nor in their syntactic relations.

Generally, intonation modelling is dealt with at two levels of description: an abstract and a concrete level. At the abstract level, attempts are made to describe human perception of intonation phenomena symbolically. The perceived intonation phenomena do not necessarily correlate with acoustic signals (*t'Hart and Cohen, 1972*). Rather, it can be best accounted for through a linguistic rendering of what is perceived. The linguistic rendering is usually based on a family of resembling pitch pattern derived from a common underlying  $f_0$  contour.

At the concrete level, on the other hand, intonation can be described in terms of those properties of the acoustic speech signal which give rise to the perceived intonation. This is usually characterised by the pattern of the change in fundamental frequency ( $f_0$ ) over time. Alternatively, the concrete level can be described by using physiological variables which control the frequency of glottal fold vibration (*Fujisaki and Hirose, 1984; Kochanski and Shih, 2003; Fujisaki et al., 2004*), for example the activities of the vocal folds.

The aim of intonation modelling, in the context of text-to-speech synthesis, is to

create a system that can predict an abstract description of intonation from an input text. The predicted abstract description is subsequently used to generate the concrete  $f_0$  contour and synthesise speech sound.

Intonation modelling in the context of speech synthesis has typically followed either a machine learning (e.g., *Wang et al. (2002); Sakurai et al. (2003)*) or theoretically motivated rule based (e.g., *Lee et al. (1993); Raptis and Carayannis (1997)*) approach. In a machine learning approach, techniques such as Artificial Neural Networks, Hidden Markov Models or Decision Trees are used to predict the intonation pattern, based on information that can be deduced from the text. While this technique can be successful if sufficient data is available, the resulting models are typically opaque, and often cannot usefully be applied outside the specific conditions under which they were trained.

In rule based intonation modelling, a set of rules is designed to transform underlying tones into an  $f_0$  contour. This often involves rules that convert discrete phonological symbols into phonetic symbols, then a final interpolation step to produce  $f_0$  contour. If the resulting set of rules is simple, this approach has the advantage that it can be understood and also transported from one speaker or situation to another.

In the following subsection, we summarise the literature on these intonation modelling approaches and discuss their strengths and weaknesses in the context of SY prosody modelling.

## 4.1 Data-driven approach to intonation and prosody modelling

In the data-driven approaches to intonation and prosody modelling, the intonation and prosodic properties of speech utterances are learned from data using automatic machine learning techniques such as *Hidden Markov Models (HMM)* (*Jensen et al., 1994; Tokuda et al., 2002a*), *Genetic Algorithms (GA)* (*Bird, 1994*) and *Artificial Neural Network (ANN)* (*Vainio, 2001*).

The philosophy underlying the data-driven model of intonation is that it is possible to build a system that captures all the prosody attributes of speech signals by automatically learning the pattern from a large speech corpus. The focus, therefore, is



on training a statistical or probabilistic model using a large annotated speech corpus. The corpus annotation may include tags for intonation events such as major and minor intonation breaks, and prominence levels of words or syllables, etc. Often the tags for labelling the prosodic events are based on phonological theory such as Tone Break and Index (ToBI) (*Wightman and Ostendorf, 1994*). After training, the intonation model is expected to generate speech intonation automatically given an input text. In essence, the intonation model learns the speech production knowledge by exploring the statistical relationships between labels of prosody information in pre-recorded speech and/or text corpus.

In the data-driven approach, the input text and associated speech are assumed to encode all the relevant prosody information. The function of the model, therefore, is to decode the encoded intonation pattern and represent them in its internal memory. In the following subsections, we summarise the literature on the techniques commonly used in data driven approach to intonation and prosody modelling.

#### 4.1.1 Hidden Markov Models (HMM) based intonation models

The motivation for using HMM in speech synthesis is its apparent successful application in automatic speech recognition (*Huang et al., 1990*). The appropriateness of using HMMs for modelling word-level prosodic and intonation information is addressed in *Ljolje and Fallside (1986)*. *Jensen et al. (1994)* presented an HMM based intonation modelling technique. The model has four basic components:

1. An acoustic front-end analyser which extracts observation vectors from the speech signal.
2. A set of HMMs representing the defined intonation units.
3. A grammar prescribing allowable sequences of intonation units.
4. A connected symbol recognition algorithm which finds the most likely sequence of intonation units given the acoustic evidence, the models and the grammar.

There are two limitations in this approach:

- it is speaker dependent,
- it assumes that an utterance consists of one and only one intonation group.

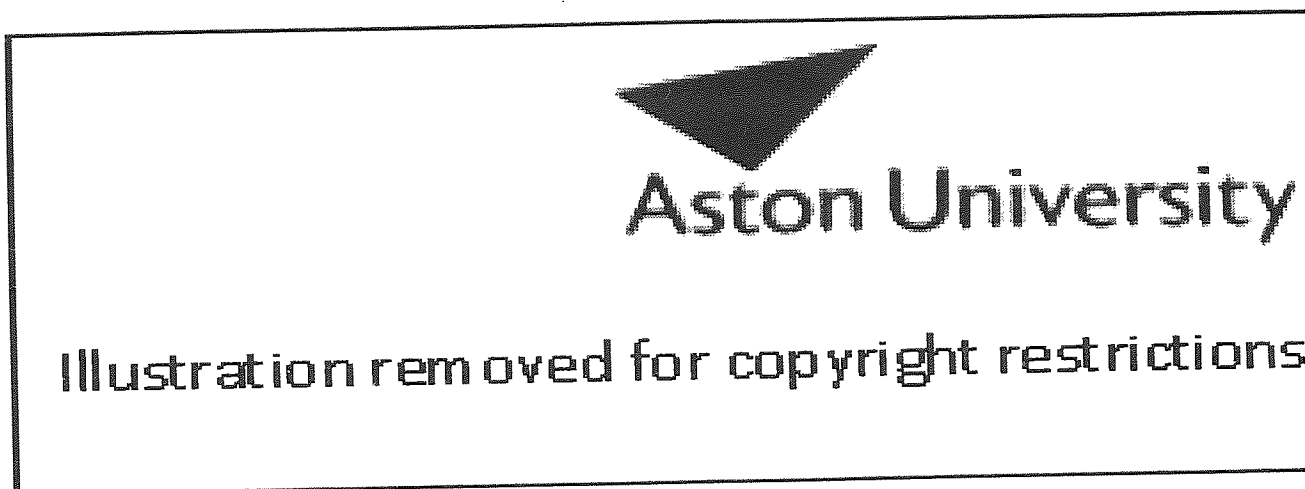


Figure 4.1: HMM based prosody model for TTS (*Tokuda et al. (2002b)*)

In more modern applications of HMM to intonation and prosody modelling, voice characteristics of synthetic speech are changed by transforming HMM parameters appropriately. For example, *Tokuda et al. (2000)* proposed parameter generation algorithms for HMM-based speech synthesis, and constructed a speech synthesis system based on this idea (see Figure 4.1). They showed that it is possible to change voice characteristics of synthetic speech by applying a speaker adaptation technique or a speaker interpolation technique. The main feature of the system is the use of dynamic features: by inclusion of dynamic coefficients in the feature vector, the dynamic coefficients of the speech parameter sequence generated in synthesis are constrained to be realistic, as defined by the parameters of the HMMs. They found that the approach can be applied successfully to Japanese  $f_0$  and duration dimensions of speech prosody (*Yoshimura et al., 1999; Tokuda et al., 2002a,b*).

*Donovan (1996)* built an HMM-based speech synthesis system in which all the system parameters used in the model were obtained by training. The cluster states of a set of decision-tree HMMs are used as the synthesis unit. The synthesis parameters for each of the clustered states were obtained through training on a single English

language speaker's continuous-speech database. This results in a very flexible system. It was reported that the model produced synthetic speech with good intelligibility and naturalness, although it sounded monotone. The HMM-based approach has also been demonstrated for German and Czech speech synthesis (*Matousek et al.*, 2002) and it is also used in Whistler, the Microsoft TTS system (*Huang et al.*, 1997).

Despite its powerful statistical approach, the application of HMM to tone-language speech synthesis is unlikely to obtain results comparable to those from non-tone languages, as demonstrated for English (*Donovan*, 2003), and accented language, such as Japanese (*Tokuda et al.*, 2000). There are a number of reasons for this. First, HMMs do not encode timing information in a way that would allow the generation of one tone per syllable (*Bird*, 1994).

Second, there is no principled upper bound on the amount of context that needs to be inspected in order to resolve ambiguities. This leads to a multiplication in the number of information states required by the HMM. This makes the training of the HMM more complicated and difficult to implement accurately. The transitions into, and out-of, a given phonetic segment with carryover or anticipatory effects due to the presence of previous or following syllable cannot be accurately accounted for in an HMM based model. This is because the same section of a pitch contour may correspond to either High (H) or Low (L) tone. For example, an H between two Hs may appear like an L tone.

Another problem is that there are a lot of segmental and super-segmental phenomena that need to be accounted for in tone language intonation. Apart from the fact that these phenomena are tone dependent, most of the parameters and features that define the phenomena overlap and are interrelated. The mismatches between hand-labelled transcriptions and HMM alignment labels can lead to discontinuities in the synthetic speech.

We consider HMM not suitable for intonation modelling in our prosody model because it will only produce decontextualised intonation models which are intrinsically unsuitable for tone-language speech synthesis. Despite its apparent success in modelling intonation for non-tone languages, these limitations probably account for the reason why there is no literature on the application of HMM to tone-language speech synthesis.

### 4.1.2 Genetic Algorithms based intonation models

In the light of the limitations of HMMs in intonation modelling, *Bird* (1994) experimented with Genetic Algorithms (GA). The aim of Bird's approach was to demonstrate the importance of having a computational tool which allows phonologists to experiment with  $f_0$  scaling of parameter.

Bird provided a parameterised  $f_0$  prediction function which generates  $f_0$  values from a tone sequence. Transcribing a sequence of  $X$  pitches (i.e.  $f_0$ ) values was considered to amount to finding a tone sequence  $T$  such that  $P(T) \approx X$ . This was modelled as a combinatorial optimisation problem and the GA algorithm was used for modelling and transcribing the intonation patterns. The gene in the GA algorithm is defined in terms of the parameters of tones in the target language. The evaluation of fitness of a gene is dependent on other genes in that population as well as the fitness of the genes in the previous ten generations. The standard GA techniques for producing a new generations, e.g. crossover, breeding and mutation, are used. The generation of new population of genes is repeated until the optimal sequence of genes is generated.

Two issues discourage the use of GA in our prosody modelling. One of them is that the methodology for the representation of tone and intonation phenomena using genes is rather unclear. The second issue is that the performance of the GA model is heavily dependent on the setting of several parameters. Finding a combination of optimal parameter setting requires a trial-and-error approach which will not necessarily produce a good model. Moreover, this approach is rarely used in the literature.

### 4.1.3 Neural Networks based intonation models

A number of studies applying Artificial Neural Networks (ANN) to speech synthesis have appeared. Just as it is for HMM-based models, the use of ANN in speech synthesis is motivated by its successes in automatic speech recognition (*Delmonte*, 2000). ANNs provide good solutions for problems involving strong non-linearity between input and output features, and also when the quantitative mechanism of the mapping is not well understood.

The basic problem in constructing and using an ANN-based model is finding an optimal network configuration and data representation as well as a method for training

the network to successfully learn intonation and other related prosody phenomena. *Vainio* (2001) experimented with a number of neural network configurations in modelling prosody for Finnish TTS. The ANNs were used to predict continuous values for fundamental frequency, loudness and segmental duration. The model assumes the availability of a text processing system that is capable of providing information about the linguistic structure of the input text in a form suitable for input to the ANN. As already discussed in the case of the HMM approach, the generation of such linguistic information requires the annotation of a speech database and its corresponding text corpus, which is a complicated and an error-prone process.

In addition, ANNs are known to converge to a local optimum solution while there are possibly better global optimum solutions. Interpreted in the context of intonation modelling, the local processing that is performed on a syllable does not give the global picture of the prosodic events of the utterance within which the syllable occurs. Furthermore, the direct mapping from linguistic to phonetic features reduces the amount of control and transparency of the model. Another problem with the ANN approach is the difficulty involved in relating the underlying mapping between inputs and outputs, expressed by the combination of weights, with specific prosodic events. This makes the performance of the model difficult to interpret, hence making incremental improvement difficult to achieve.

To circumvent these problems, a hybrid approach which integrates ANN and HMM have been suggested by *Hendessi et al.* (2002). However, the most promising results are those employing Recurrent Neural Networks (RNN). For example, *Chen et al.* (1998) proposed a four-layer Recurrent Neural Networks (RNN) for synthesising Mandarin Chinese prosody. The input layer and the first hidden layer of the RNN operates with a word-synchronised clock. This is done to account for current-word phonological states within the prosodic structure of the text to be synthesised. The idea underlying this approach is that it is possible to use a model to explore the relationship between prosodic phrase structure and the linguistic features of input text for simulating human's prosody mechanism.

To make the model more robust, *Chen et al.* (1998) incorporate additional syllable-level data that are fed directly into the second layer of the RNN. This allows the second



Figure 4.2: RNN based prosody model (*Wang et al.* (2002))

hidden layer and the output layer to operate on a syllable-synchronised clock and use outputs from the preceding layers, along with additional syllable-level input fed directly to the second layer, to generate the desired prosody parameters.

An improved version of the RNN based model was later developed by *Wang et al.* (2002), also for Mandarin prosody modelling. In order to account for a wider context in the generated synthetic prosody, feature sets of the current syllable segment and its six nearest neighbours were assembled and inputted into the RNN (see Figure 4.2). *Wang et al.*'s model was shown to possess a good ability to learn the complex relationship of the input feature vector sequence and the output target by implicitly storing the contextual information of the input sequence in its hidden layer. So the method is suitable for the problem of realising a complex mapping between the input prosodic features and the output linguistic features.

*Sakurai et al.* (2003) experimented with three types of neural networks for text-to-speech synthesis of Japanese: the multi-layer perceptron (*MLP*), *Jordan* (a structure having feedbacks from output elements) and *Elman* (a structure having feedbacks from a hidden layer). *Jordan* and *Elman* are partial recurrent neural networks. They found that their models outperformed some rule-based models both in qualitative and

quantitative evaluation.

There are a number of problems that render RNN unsuitable for prosody modelling in the context of tone language TTS. First, although the phonological rules of tone modification may be implicitly learned and stored in the weights of the RNN (*Lin et al.*, 2003), they cannot be explicitly extracted from an RNN-based prosodic model. Noting this limitation, *Lin et al.* (2003) proposed a Recurrent Fuzzy Neural Network (RFNN) for Mandarin speech synthesis. The aim was for the model to mimic the experience of a human expert as a speech knowledge-based system where the knowledge is stored in a set of fuzzy IF-THEN rules.

The main weakness of this approach is that it still suffers from the problem inherent in RNN-based models in that the knowledge that can be represented is constrained by the RNN component of the model. In addition, this approach presupposes that human linguistic behaviour is adequately represented in written text and can be learned automatically by machines. Several findings have shown that this is not the case (e.g. *Quené and Kager* (1992); *Monaghan* (1993); *Prevost and Steedman* (1994); *Olaszy and Németh* (1997); *Greenberg* (1998); *Möbius* (2003); *Smith* (2004)).

## 4.2 Theoretical models of intonation and prosody

The development and paradigms underlying most modern intonation models used in TTS applications is outlined in *Botinis et al.* (2001). All the models are based on the AM or the HO approach (cf. Section 1.3.1), or various degrees of their mixture. The aim here is to discuss the literature on intonation modelling in the context of tone language prosody with a view to highlight their strengths and weaknesses as a possible candidate for modelling SY prosody.

In this work, we will not be using the intonation models proposed for non-tone language. In this section, however, we will review the important works in both tone and non-tone language intonation and prosody modelling. The aim of the review is to highlight the important strengths and weaknesses of the models used in speech synthesis as well as how our model is developed based on some ideas borrowed from these models.

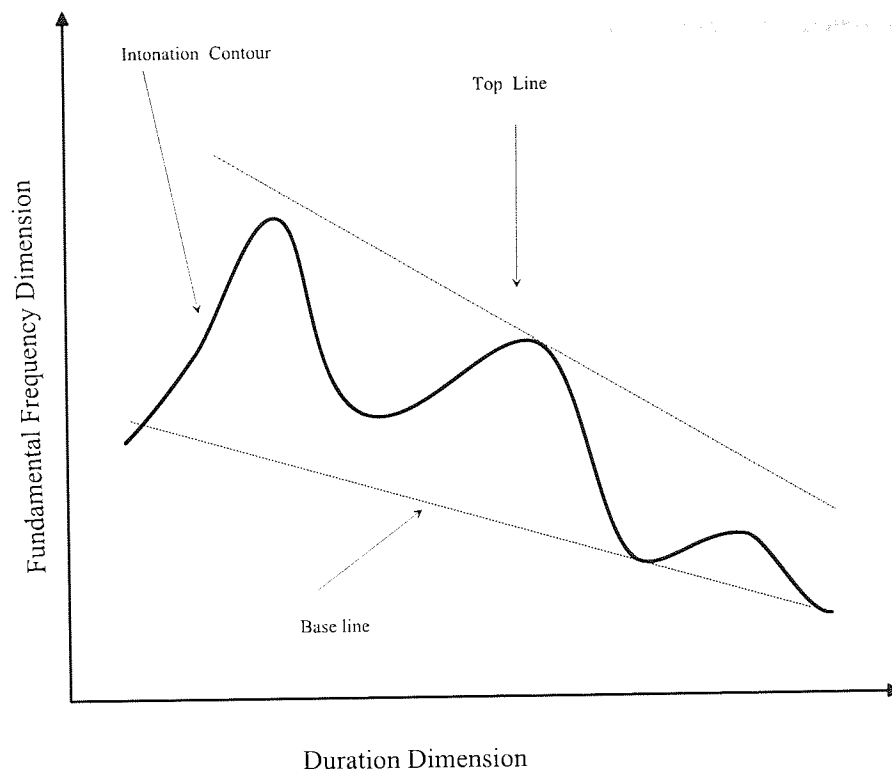


Figure 4.3: Top line and Bottom line in the Pierrehumbert model

#### 4.2.1 Pierrehumbert's model of intonation

Pierrehumbert's intonation model is built around metrical phonology (*Lieberman and Prince, 1977*) and auto-segmental phonology (*Leben, 1976*). In this model, an intonation phrase is represented as a sequence of high (H) and low (L) tones. Although the tones follow each other sequentially, they are members of a primary phonological opposing tones which do not interact. To generate the  $f_0$  contour of an utterance, the prosodic contribution to the  $f_0$  is analysed as a series of target values within the current  $f_0$  range. This range is normally indicated by two lines known as the top line and bottom line (see Figure 4.3).  $f_0$  targets and their alignment with text are indicated by decimal numbers, which are interpreted as a proportion of the way from the bottom to top of the  $f_0$  range.

*Lindau (1986)* applied this model to the synthesis of Hausa, a two-toned African language, using the following algorithm:

- Set the duration for the given number of syllables.
- Set the value for the first high.



- Construct the intonation grid, anchored on an arbitrary (speaker-specific) value for the first high. This is done by specifying the basic top line and bottom line slopes.
- Insert low start and end points.
- Insert high and low tones on the grid lines. The highs and lows will constitute turning points in the intonation curve. Align the sentence so that the highs and lows are syllable boundaries.
- Apply local tone assimilation rules. This will result in a steeper slope, the more low-high sequence the sentence has. Optional rules may be applied that move specific tones away from the grid lines.
- Concatenate all the turning points into a smooth curve. The smooth trajectory is generated by interpolating between turning points, using a piecewise application of third degree polynomials. Each turning point is specified as  $t_n, y_n$ , where  $y$  = fundamental frequency at time  $t$ .

The input to Lindau's model is a phonetic transcription of sentence with lexical tones marked. Lindau found that the rules and the interpolation in the above procedure approximate the data well. However, the perceptual evaluation of the model was not reported making it difficult to determine its suitability for TTS application. It is well known that, accurate data fit is not enough for perceptual quality of synthetic speech prosody ('tHart, 1991).

The *Tone and Break Indices* (ToBI) scheme is a linguistic category annotation scheme developed around the Pierrehumbert's intonation model and widely used in TTS. One reason for the widespread application of ToBI was the growing emphasis on statistical data-driven methods, which were driving speech synthesis and related technologies. The data-driven methods require the automated analysis of large speech corpus. The application of automated analysis tools, however, requires the use of a standard annotation system which can be used to annotate the speech corpus in a consistent manner. The ability of ToBI to fulfil this requirement led to its widespread applications.

The ToBI system has been developed for a number of languages including: German (*Grice et al.*, 1996), Japanese (*Campbell and Venditti*, 1995; *Campbell*, 1997), and Korean (*Beckman and Jun*, 1996). The main similarity among these systems is that break indices and tones together define two levels of phrasing, minor or intermediate

and major or intonational, with the former delimited by a phrase accent and a boundary tone.

The ToBI model for American English (AME-ToBI) is described by *Lieberman and Sproat* (1992) and *Pitrelli et al.* (1994). The AME-ToBI system consists of annotations at four or more time-linked levels of analysis. The three obligatory tiers are: (i) an orthographic tier of time-aligned words; (ii) a break index tier indicating degree of junction between words, its values ranges from 0 ‘no word boundary’ to 4 ‘full intonational phrase boundary’; and (iii) a tonal tier, where pitch accents, phrase accents and boundary tones describing targets in the  $f_0$ , defines intonation contours. A fourth tier, the miscellaneous tier, is provided for any additional phenomena that may be required for specific implementation of the model.

In the AmE-ToBI described by *Syrdal et al.* (2001), a minor phrase accent consists of a word, or strings of words, with at least one pitch accent aligned with the rhythmically strongest syllable of the accented lexical item(s). This is followed by a phrase accent which may be high (**H-**) or low (**L-**). The major phrases consist of one or more minor phrases, plus a high or low boundary tone (**H%** or **L%**) at the right end of the phrase. A starred tone, i.e.  $H^*$  or  $L^*$ , indicates which tone is aligned with the stressed syllable of the word bearing a complex accent. For example, a standard declarative pitch contour will end in a low phrase accent and low boundary tone and is represented by **L-L%**; a standard yes-no question contour ends in **H-H%**.

Downstep is a compression and lowering of pitch range, and is effectively what makes the Pierrehumbert theory work with only two tones, i.e. H and L (*Clark*, 2003). Pierrehumbert considers downstep to be triggered by an H L H tonal sequence which includes a bi-tonal pitch accent such as in the sequence  $H^*+L H^*+L L-L\%$ . The result of downstepping is that the second  $H^*$  is realised at a lower than expected  $f_0$  value indicating that the pitch range has been lowered and compressed. The notation ‘!’ is attached to accent descriptions to show they are downstepped. For example,  $H^*$  becomes  $!H^*$  when downstepped,  $L^*+H$  becomes  $!L^*+H$  and so forth. *Final lowering* is an intonation phenomena in which extra pitch lowering occurs at the end of phrases or utterances, thus making the pitch reach a lower level than elsewhere in the utterance.  *$f_0$  reset*, is an intonation phenomena whereby  $f_0$  jumps back to some higher level after

reaching a low point.

It is important to note that *Pierrehumbert's* (2000) model of downstepping has been questioned in *Ladd* (2000) where it was suggested that downstep is controlled by an independent feature which can be set on given pitch accents. The extension of the definition of 'tone' and 'association' in *Pierrehumbert* model to tone language intonation modelling has also been questioned by *Ladd* (2000).

This approach may, therefore, not produce an accurate account of the realised intonation phenomena because the same concept is used to describe other very important intonation phenomena in SY, i.e. final lowering and  $f_0$  reset. In fact, *Wightman* (2002) observed that the original design goal for providing a description of intonation contour has compromised the ToBI system. For example, *Wightman and Ostendorf* (1994) found that ToBI labelling relies on linguistic judgements made by experts and is consequently difficult to carry out automatic labelling. We conclude with *Wightman's* (2002) argument that: "*the need for the descriptive component of ToBI no longer exist and that we should only be labelling what we hear: our perception*".

#### 4.2.2 Taylor's model

*Taylor* (2000b) presents an intonation model, called the *Tilt* model, based solely on the acoustic details of speech stream. The *Tilt* intonation model comprises four basic intonation units: pitch accents, boundary tones, connection, and silence. Every event contains a description for both the rise and the fall. Each of the descriptions includes the  $f_0$  at the start of the unit. Pitch accents and boundary tones are also described by their duration, absolute amplitude (the sum of the  $f_0$  movement over the event), the position at which the rising portion of the event stops and the fall begins (i.e. peak) and the *Tilt* value.

A *Tilt* labelling for an utterance consists of an assignment of one of the four basic intonation events discussed above together with a number of continuous parameters. The *Tilt* parameter represents the amount of fall and rise in the accent. The starting  $f_0$  of an accent acts as a point from which all other calculations may be made. The absolute amplitude for the starting  $f_0$ , i.e.  $A_{rise}$ , of the peak is the first portion of the absolute amplitude parameter. The other portion is the absolute amplitude parameter

from the peak to the end of the event,  $A_{fall}$  (see Figure 4.4). The two amplitudes are added together to form the absolute amplitude value. The Tilt parameter is the difference of the amplitudes divided by their sum. This is computed as shown in Equation 6.3.

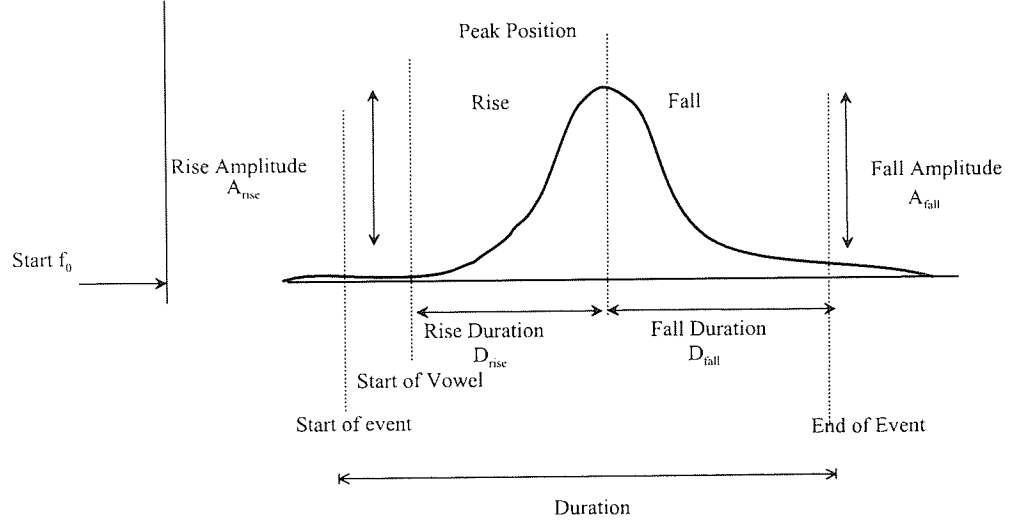


Figure 4.4: Graphical illustration of the *Tilt* model

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \quad (4.1)$$

Similar computation is performed for the duration dimension. In this case, the corresponding durations  $D_{rise}$  and  $D_{fall}$  are used to compute the Tilt value for the duration dimension as follows:

$$tilt_{dur} = \frac{|D_{rise}| - |D_{fall}|}{|D_{rise}| + |D_{fall}|} \quad (4.2)$$

These parameters are then converted into three Tilt parameters:  $Tilt$ ,  $A_{event}$  and  $D_{event}$ . The parameters are computed as follows:

$$tilt = \frac{1}{2}tilt_{amp} + \frac{1}{2}tilt_{dur} \quad (4.3)$$

$$= \frac{|A_{rise}| - |A_{fall}|}{2 \times (|A_{rise}| + |A_{fall}|)} + \frac{|D_{rise}| - |D_{fall}|}{2 \times (|D_{rise}| + |D_{fall}|)}$$

$$A_{event} = |A_{rise}| + |A_{fall}| \quad (4.4)$$

$$D_{event} = |D_{rise}| + |D_{fall}| \quad (4.5)$$

The Tilt value is a dimensionless parameter in the range  $[-1, 1]$ , and describes the shape of an intonation event. A value of -1 indicates a pure fall, 1 denotes a pure rise, and 0 indicates that the event contains equal options of rise and fall. Two other parameters of the Tilt model are:

**Position:** the peak location of an intonation event, which is usually defined as the distance between the vowel starting time and the peak location.

**Start  $f_0$ :** the absolute  $f_0$  at the start of an intonation event.

In order to realise  $f_0$  values from the above phonetic representation of intonation a data-driven approach is required. An example of the application of such data-driven approach is the Hidden Markov Model (HMM) technique described by *Taylor (2000b)* for an intonation event detector based on Tilt. In another work, *Dusterhoff et al. (1999)* presented an intonation generation system which uses classification and regression trees to predict the location and fundamental parameter of intonation events within the Tilt intonation model. They reported an overall accuracy of 82.62%.

The Tilt Model and AmE-ToBI model are similar in two respects (*Taylor, 2000b, 1992*): (1) both models are based on AM theory in that intonation is modelled as a linear sequence of events based on intonational entities rather than as a superposition organisation, (2) both models implement the idea of an abstract phonological representation of intonation.

Two problems make the Tilt model unsuitable for our purpose. The first, and most significant problem, is that there is no explicit way to link the intonation patterns with the tonal description. It is, therefore, impossible to determine which syllable in an utterance intonation is carrying a particular tonal signature. This is because the Tilt model treats pitch accent and tones as events characterised by continuous acoustic-phonetic parameter. This approach has been criticised as being paralinguistic (*Ladd, 1996*). The second problem is that the model is too low-level for direct phonological analysis.

## 4.2.3 INTSINT

International Transcription System for INTonation (INTSINT) is a phonological intonation description system in which intonation is described with a limited set of abstract tonal symbols (*Campione et al.*, 1997). The system is designed in a way that separate inventories of pitch patterns for different languages are not required (*Campione et al.*, 2000). The abstract symbols defined for representing target points are:

- T- Top
- M- Mid
- B- Bottom
- H- Higher
- S- Same
- L- Lower
- U- Up-stepped
- D- Down-stepped

The symbols T, M and B represent absolute abstract tones and they model a speaker's overall pitch range. The symbols H, S, L, U and D are relative tones and they are defined relative to the values of the preceding target points. These relative tones are further divided into two groups. The non-iterative group, which includes H, S and L tones, cannot occur repeatedly. The iterative group, i.e. U and D tones, can occur repeatedly. The model assumes that iterative rising or lowering uses smaller  $f_0$  intervals than their non-iterative counterparts. Table 4.1 shows the orthographic code and icons used in the INSTINT coding system.

The input to INTSINT is a series of target points, which is estimated from acoustic low-level modelling techniques such as MOMEL (MOdélisation de MELodies) or by using automatic coding such as CART (*Campione et al.*, 1997; *Louw and Barnard*, 2002). *Campione et al.* (2000) demonstrated six different implementations of INTSINT based on MOMEL. In all the implementations, two absolute symbols, T and B, are used to code extreme pitch values. The symbol S is used to code target points which are not significantly different from the preceding point. Targets higher than a threshold  $\tau_T$  are coded as T, those below a threshold  $\tau_B$  are coded as B. Target points less than 2.5%

Table 4.1: INTSINT coding system

		Positive	Neutral	Negative
ABSOLUTE		T [ $\uparrow$ ]	M [ $\Rightarrow$ ]	B [ $\Downarrow$ ]
RELATIVE	Non-Iterative	H [ $\uparrow$ ]	S [ $\rightarrow$ ]	L [ $\downarrow$ ]
	Iterative	U [ $<$ ]	•	D [ $>$ ]

(on log scale) from the preceding point are coded S. The target  $\tau_T$  and  $\tau_B$  are chosen so that 5% of the target points are coded T and another 5% are coded B, assuming a normal distribution of the values of the target points.

The MOMEL (*MOdélisation de MELodic*) algorithm gives a representation of the melodic curve, which characterises the temporal variations of the laryngeal frequency, by the way of a quadratic spline function. MOMEL is a low-level  $f_0$  modelling technique capable of analysing and synthesising  $f_0$  curves automatically. The MOMEL technique generates  $f_0$  in four steps. In the first step, the raw  $f_0$  values extracted from acoustic signal are pre-processed. This pre-processing has the effect of eliminating erratic  $f_0$  values. In the second step, target candidates are estimated. The purpose of this is to obtain the most suitable quadratic function for representing the target points.

In the third step, the sequence of target candidates are partitioned by using a moving window of length about 300ms to cover the acoustic signal. The target in the first half of the window is compared to the average value in the second half. In each window, the  $f_0$  curve is calculated, and an approximation can be given by such a polynomial, with the only purpose to minimise the quadratic error between the initial curve and the polynomial. In the final step, points that are more than 5% below the polynomial are set to zero. A polynomial is then recalculated with the remaining points. The process is repeated until no new points are set to zero.

A major problem with INTSINT is that of tonal alignment in the generated intonation contour. This problem arises due to the tonal symbols corresponding to the curves calculated from phonetic data generated by MOMEL. This implies that there is no direct connection between the INTSINT labels and phonological data. Thus the INTSINT label may not be time-aligned with some of the linguistic features that can be extracted from the orthographic data. An accurate time alignment of  $f_0$  contour with

segmental string is very important for the intelligibility and naturalness of synthesis.

A major weakness of MOMEL is that it does not provide a relative scaling of target points on the  $f_0$  scale. In the context of our prosody modelling approach what is important is generally not the absolute  $f_0$  as described by the model but the relative value. For example, the peak  $f_0$  value perceived as an H tone in one context may, in fact, be perceived as an L tone in another context. These weaknesses limit the application of the model in tone language TTS because we will not be able to exploit the tonal information and effectively implement intonation contour based on the language orthographic data.

#### 4.2.4 Fujisaki model

*Fujisaki and Hirose* (1982) proposed a physiologically motivated and an hierarchically structured quantitative model of intonation which has been used for the analysis and synthesis of complex  $f_0$  contours. The model additively superposes or overlay a basic  $f_0$  value ( $F_{min}$ ), a phrase component, and an accent or tone component on a logarithmic scale (see Figure 4.5). The phrase component of the model represents the global slope and the slow variations of the  $f_0$  contour in the utterance. The phrase component can be used to implement  $f_0$  declination since the phrase contour reaches its maximum rather early and descends monotonically along the major part of the utterance.

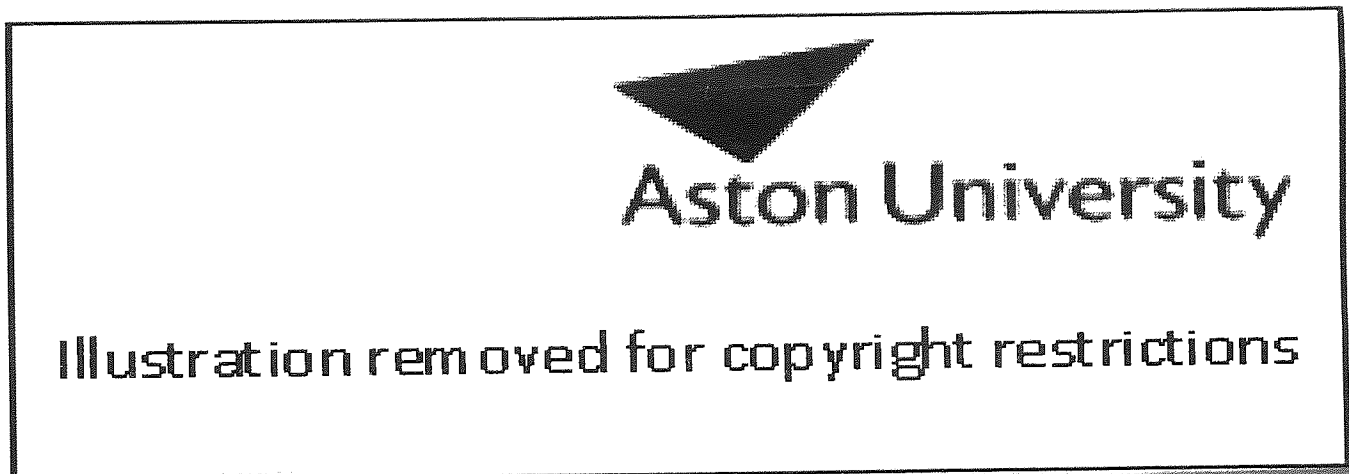


Figure 4.5: The structure of Fujisaki model (*Clark* (2003))

The declination phenomenon can be realised if the contour which results from adding basic value  $F_{min}$  and the phrase component can be interpreted as the baseline



of the intonation contour. The local, more rapidly changing aspects of the intonation contour are represented by the accent or tone component. These  $f_0$  movements are superimposed onto the global contour and can be related to the realisation of tones on tone language syllables. The accent component is made up of partial contours that are in turn generated by accent commands. Each accent group (*Möbius et al.*, 1993) is modelled by the contour resulting from exactly one accent command.

The control mechanisms of the two components are realised as critically damped second-order system responding to functions. The phrase component responds to impulse functions and the accent component responds to rectangular functions. These functions are generated by two different sets of parameters. One of them is the timing and amplitudes of the phrase commands as well as the damping factors of the phrase control mechanism. The other is the amplitudes and the timing of the onsets and offsets of the accent commands as well as the damping factors of the accent control mechanism. All these parameter values are constant for a defined time interval.

The Fujisaki model can be represented mathematically by Equation 4.6 (*Fujisaki et al.*, 2005):



Aston University

Content has been removed due to copyright restrictions



Aston University

Content has been removed due to copyright restrictions

$\theta$  a parameter to indicate the ceiling level of the tone.

The Fujisaki model can be characterised as a functional model of the production apparatus. The model represents each partial glottal mechanism of fundamental frequency control as separate component. Using an analysis-by-synthesis procedure, the complex  $f_0$  contour of a given speech utterance is decomposed into the components of the model. For example, the time constant,  $\alpha$ , governs how quickly the phrase reaches its maximum value, and how quickly it falls off after this. The accent time constant, on the other hand, is usually much higher than  $\alpha_i$  and it gives the filter a quicker response time. This implies that the shape produced from the accent or tone component reaches its maximum value and falls off much faster than the phrase component. The pitch range can be raised by increasing the amplitude of the phrase component. The  $f_0$  peak of tones can be varied by changing the amplitude of a step input. Using these information, it is possible to obtain an accurate approximation of the original  $f_0$  contour by successively optimising the parameter values.

The Fujisaki model has been applied in the modelling of the Tokyo dialect of Japanese declarative sentences (*Fujisaki and Hirose, 1984*), Mandarin Chinese (*Fujisaki et al., 2000*), English and Estonian (*Fujisaki et al., 1998*), and German (*Möbius et al., 1993*). A major advantage of the model is its flexibility since the degree of freedom of

its parameter can, in principle, approximate any given intonation contour. Also, the model requires a small number of parameters to represent an  $f_0$  contour.

The major weakness of the model is that it has a physiological motivation, hence it does not come with a set of formula for describing how to relate its parameter to linguistic units and structure. Although the accent and phrase components have their linguistic correspondence in Japanese, these do not coincide with similar structure in tone-languages. This implies that a linguistically feasible interpretation must be developed for each individual language. In addition, the model parameter, i.e. phrase and accent commands, must be derived from the natural  $f_0$  contour using a manual and error-prone procedure. Reasonable algorithms for doing this derivation have been developed (*Mixdorff*, 2000), but they require a high degree of expert knowledge and post-processing to work efficiently in a practical TTS system.

Furthermore, prosodic phenomena which require a finer control of the  $f_0$  contour, such as H rising in SY, cannot be expressed by the model. This limits its application to SY intonation modelling. Also, the model does not include a means for incorporating perceptual information into intonation realisation. This is a crucial requirement for TTS design, evaluation and further improvement.

#### 4.2.5 The Stem-ML model

The Soft-TEMplate Mark-up Language (Stem-ML) was introduced by *Kochanski and Shih* (2003) as a prosody tagging and generation system. Stem-ML is motivated by the ideas that a quantitative model of intonation is more efficient if the  $f_0$  curve representing tonal or accent elements is structured into intonational entities that can be directly related to linguistic features. The system combines mark-up tags and pitch generation into a single process. The mark-up tags are mathematically defined and  $f_0$  generation is deterministic (*Kochanski and Shih*, 2003).

The physical modelling in Stem-ML was inspired by tone languages such as Mandarin (*Kochanski et al.*, 2003b). Isolated syllables in tone languages have pitch curves close to the ideal shapes of their tone. In sentences, however, tones interact and their shapes vary from the ideal ones due to the co-articulation phenomena (*Xu*, 1999a). The Stem-ML model follows the HO approach in that it assumes that speakers are

capable of pre-planning their pitch pattern a few syllables ahead.

Stem-ML assumes that the prosodic trajectory is continuous and is smooth over a short time scale. Underlying this assumption is the idea that all aspects of prosody are controlled by muscle actions, and that mapping between muscle activation and perception is not strongly non-linear.

There are two levels of tags defined in Stem-ML. Level 1 tags correspond to the actual Stem-ML model of intonation events and serve as the basis for  $f_0$  generation. Level 2 tags are contained in input text and are used for specifying prosody control through linguistic definitions.

There are three components in Stem-ML prosody modelling:

- A linguistic modelling component that converts Stem-ML Level 2 tags into Level 1 tags.
- A pitch generation component that takes the Stem-ML Level 1 tagged text and produces a time series of pitch values.
- An optional segmental effect component that calculates how  $f_0$  depends on the phoneme sequence.

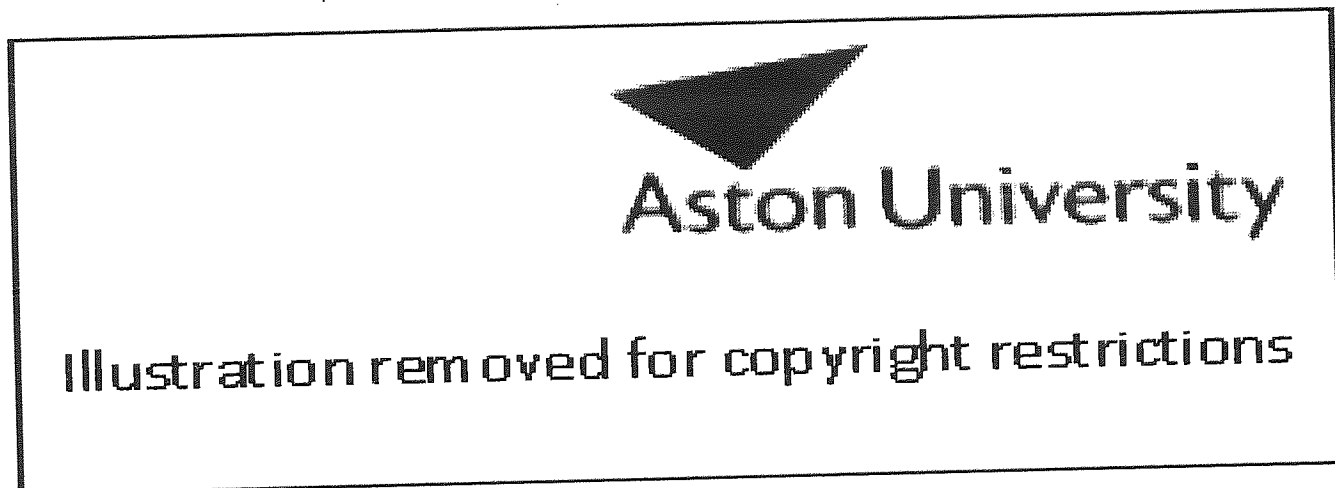


Figure 4.6: Block diagram of the Stem-ML algorithm (*Kochanski and Shih (2003)*)

Figure 4.6 is the block diagram of the Stem-ML algorithm. The steps in the algorithm are (*Kochanski and Shih, 2003*):

- calculate the phrase curve,

- calculate the prosody, relative to the phrase curve,
- map from an abstract description of prosody to observable quantities.

Internally, each tag is defined mathematically with parameter settings describing variations. The dotted line boxes show the tags that influence each step. For example, <step/> and <slope/> are two types of tags that can be used to define phrase curve, and the <stress/> tags allow users to specify tone or accent templates. Each tag puts a set of constraint on the prosody. These constraints enforce smoothness and continuity of an  $f_0$  contour. Technically, the algorithm implements a regularised fit to constraints by way of a least-mean-square solution of the constraint equations.

Because Stem-ML is defined in physiological terms and the Stem-ML tags are not associated with particular language features, it has the potential of being a language independent model of prosody. *Kochanski and Shih (2003)* found that the prosodic measurements from the model also show a useful correlation with word length and part-of-speech of words. They also found that the model show that the strength values correlate in expected ways with other acoustic observations such as duration.

A major weakness of this model is that it is not clear how the parameters of the model can be subjected to perceptual constraints. This makes it difficult to use perceptual information for improving the naturalness quality of the synthetic speech produced by the model. Moreover, the Stem-ML model relates prosody directly with the physiology of speech production in terms of articulatory movements. It is well known that the relationship between speech sound and the speech production mechanism is not a straight forward one. However, the apparent success of the Stem-ML model at modelling intonation and prosody for some tone languages has motivated us to experiment with the model. Our experiment using the Stem-ML model for modelling SY prosody is documented in Section 12.4.

#### 4.2.6 Gibbon's computational model of intonation

*Gibbon (2003b)* takes a computational approach to intonation modelling. This approach uses both symbolic and numeric computational methods in modelling intonation and prosody. *Gibbon (2004a)* argues that in order to develop a model that is consistent and precise and can be evaluated, both symbolic and numeric computation

methods are required. For the computational modelling of tone patterns, he suggested that it is necessary to define appropriate data structures and operational devices which can recognise and generate these structures.

Based on the principle of automata theory and formal grammar, *Gibbon* (2003b) showed that the sequence of tone-allotone pairs which occurs in terrace tone sequences are easily represented by a Finite State Transducer (FST). *Gibbon's* approach was motivated by the linguistic use of FST, where the relation between FSTs and standard phonological rule types has been demonstrated. The advantage of using FSTs is that the overall system is clearly illustrated, and verification of the description by exhaustive generation is supported. In contrast, a collection of isolated rules are not always perspicuous, and cannot be easily evaluated.

*Gibbon* (2003a) applied the FST modelling approach to several West African tone languages and found that terracing patterns appear as *iteration*, *cycles*, and *oscillations*. He also found that the typological difference between the languages are clearly reflected in the structuring of the FSTs used to describe them.

In the *Simplest tonal schema* (see Figure 4.7(a)), the FST has a start state, at which initial H and L tone values are defined, and two states, one for H tone and another for L tone sequences. There are language specific transitions between these states. These transitions are associated with tonal relation such as total or partial automatic downstep and upstep. The *Generalised schema* (see Figure 4.7(b)), uses a two-level system in addition with a *terrace* defined as a complete cycle which includes both H state and the L state, and a *demi-terrace* as a sequence of transition leading to the same target state.

The implementation of the model for two African tone languages, i.e. Tem and Baule, was demonstrated. In Tem (see Figure 4.7(c)), additional state was introduced in order to handle the additional complexity of the final low tone which falls to a more or less constant level. Baule has a further complexity (see Figure 4.7(d)), in that the contexts for tonal assimilations are non-adjacent. This necessitates longer path between the main states, also with a non-deterministic element.

This model associates the linguistic terminology, which is generally in use for different tonal processes, with transitions in the lexical tone automaton. The input and

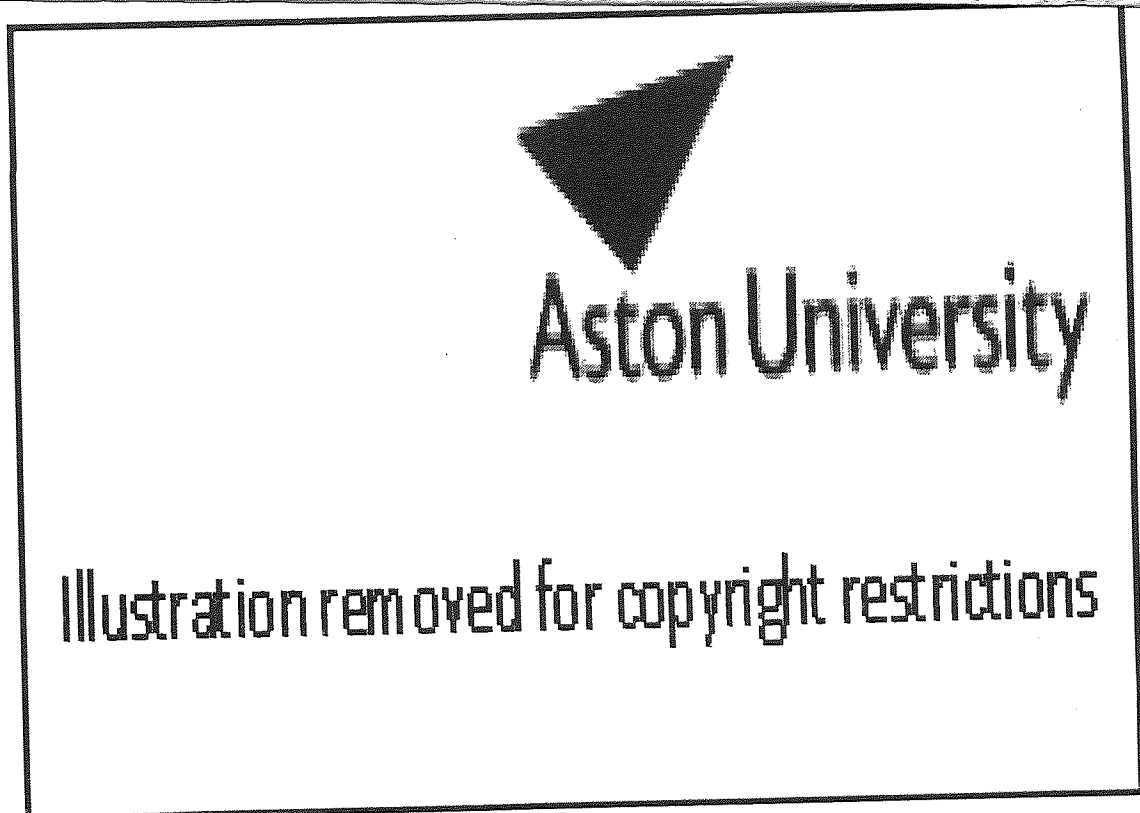
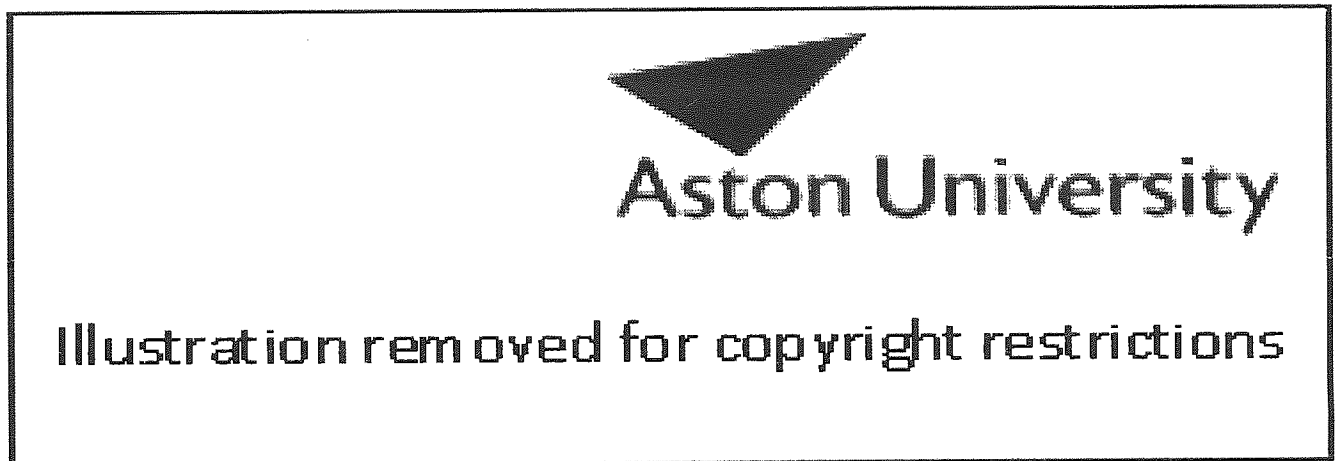


Figure 4.7: Typologically distinct lexical tones automata (*Gibbon (2004a)*)

output of the terracing FST relate to grammatical and phonetic constraints respectively.

It is possible to see the Gibbon's models as a kind of scheduler for the application of phonological rules. It is well known that phonological rules of this kind can be modelled by finite state transducers. The *Gibbon* (2004b) approach for the simplest case of a two-tone language, such as Hausa, will use a three state automaton. The automaton will have the following states:

- An initial state, `state_s`.
- One state H tone, `state_h`.
- One state L tone, `state_l`.

The states for tone H and tone L, i.e. `state_h` and `state_l`, are also final states. The six basic transitions are:

```
<state_s, H, state_h>
<state_s, L, state_l>
<state_h, H, state_h>
<state_h, L, state_l>
<state_l, L, state_l>
<state_l, H, state_h>
```

The symbols H and L in this basic automaton are extended to different transduction functions on each transition as follows:

```
<state_s, H->initial_h, state_h>
<state_s, L->initial_l, state_l>
<state_h, H->terrace_h, state_h>
<state_h, L->upstep_l, state_l>
<state_l, L->terrace_l, state_l>
<state_l, H->downstep_h, state_h>
```

The transduction functions can be regarded as symbolic, for phonological purposes, or as numerical functions, for acoustic purposes. Gibbon's methodology is similar to



that of *Lieberman and Pierrehumbert* (1984) except that an exhaustive search is used without the downhill simplex search heuristics used. The numerical functions follow *Lieberman and Pierrehumbert* (1984) on intonation, and work by *Akinlabí and Lieberman* (2000) on Yorùbá tone, and essentially generate asymptotic functions by multiplying the previous value with a typical larger or smaller value (with a baseline offset value) as discussed in *Láníran and Clements* (2003).

The problem of identifying prosodic timing hierarchies and their relation to syntactic hierarchies was addressed using a two stage process. In the first stage, an automatic timing tree induction from local duration differences in annotated speech signal data is performed. In the second stage, the automatic calculation of a Tree Similarity Index (TSI) between the resulting timing trees and grammatical trees is computed. A possible architecture for the application of this model to TTS system implementation is shown in Figure 4.8.

It is possible to implement this model for a three tone language, such as SY, but the process is more complex. An advantage of this approach is that it is generic. Its use of Finite State Transducers (FST) makes it easy to implement. This is because FSTs are simple mechanisms and there are readily available tools for implementing them. We have not used this approach in our intonation modelling because it is not clear how we can integrate perceptual information into the model. Also, the model treats intonation phenomena such as downstep and upstep as discrete entities. However, the domain of such phenomena are somewhat fuzzy and requires a compositional model which integrates information at the phonological (i.e. discrete), phonetic and acoustic levels of prosody.

#### 4.2.7 Ladd's intonation model

Underlying Ladd's intonation model is the idea that  $f_0$  contours can be modelled as sequences of abstract phonological elements (tonal configurations and the register steps) inserted into a text string (*Ladd*, 1987). In this model,  $f_0$  specification is seen as a process involving two major stages: (i) a text-to-phonology stage in which the text string is annotated with indications of "intonational events"; and (ii) a phonology-to-phonetics stage, which converts the abstract elements of the phonological strings into

Figure 4.8: Gibbon architecture for prosody modelling (*Gibbon (2004a)*)

a sequence of  $f_0$  targets aligned with the segmental sounds. It also generates the  $f_0$  transition that links the targets in the output contour.

Specifically, the text-to-phonology stage builds up an abstract representation of the contour in terms of two kinds of phonological elements: register steps and tonal configurations. The phonology-to-phonetic stage defines an invariant speaker range in terms of speaker specific base-line ( $F_{min}$ ) and a default sentence-initial register setting ( $N$ ). Any given register setting  $f(N)$  is defined by the equation:

$$f(N) = N \times d^i \quad (4.9)$$

where  $d$  is the step size and  $i$  is an integer. The top of the register (i.e. High tone) at any given point is  $f(N) \times w$ , where  $w$  is a parameter determining the width of the register, and the bottom of the register (i.e. Low tone) is  $F(N)/w$ . In this model, tonal configurations are defined as sequence of values of  $T$  in the equation:

$$f(T) = W^T \quad (4.10)$$

where  $T = +1$  for High,  $-1$  for Low, and  $0$  for middle of the register. The actual  $f_0$  values for the phonologically specified targets are then computed by the equation:

$$F0 = F_{min} \times f(N) \times f(T) \quad (4.11)$$

The model was intended as a universal model of natural speech, not just a device for producing good synthesis. To test this universality hypothesis, the model has been fitted for English and Standard Yorùbá (*Ladd, 1987*). In the synthesis of English, fixed values of  $W(1.45)$  and  $d(0.80)$  gave natural sounding results for many different values of  $F_{min}$  and  $N$ .

In the case of SY, the model gave an excellent fit “if and only if”  $W$  is allowed to be a speaker-dependent variable. This shows that the implementation of the model for different languages requires the manipulation of different model parameters. This model is similar to the *Fujisaki and Hirose (1982)* and *Lieberman and Pierrehumbert (1984)* in two respects. First, they incorporate a mechanism for modelling the local and global transition of the  $f_0$  contour. Second, they implement the AM approach to intonation modelling. An important strength of Ladd’s model, however, is that it has the potential to create different speaker characteristics. Furthermore, the model parameters and variables can be used to explain the way pitch range varies within and between speakers.

A refined version of the Ladd’s model was applied in the Aculab TTS system (*Monaghan, 1990*). The parameters of that model are shown in Figure 4.9. In this model, the parameter  $B$ , which is the speaker’s lowest pitch, specifies the global bottom line of the speaker’s  $f_0$  range. Three other tonal levels are identified: Top of the current range ( $H$ ), neutral position in current range ( $M$ ) and the bottom of the current range ( $L$ ). The current range,  $R$ , varies for different segment of an utterance. This configuration facilitates the synthesis of a wide range of different voices simply by modifying a handful of parameters. For example, the frequency value of the neutral position in the current  $f_0$  range ( $M$ ) changes as a function of the prosodic structure, responding to boundaries and to upsteps and downsteps. The top ( $H$ ) and bottom ( $L$ ) of the current range move proportionately with  $M$ . The distance between  $H$  and  $L$  varies as a function of local prominence or emphasis.

The Aculab TTS system has been used for prosody modelling for many languages including British and American English, German, French, Italian, Brazilian Portuguese, and Latin American Spanish, with multiple configuration for each language (*Monaghan, 2003*). Some aspects of the model (e.g. the lowest and highest possible

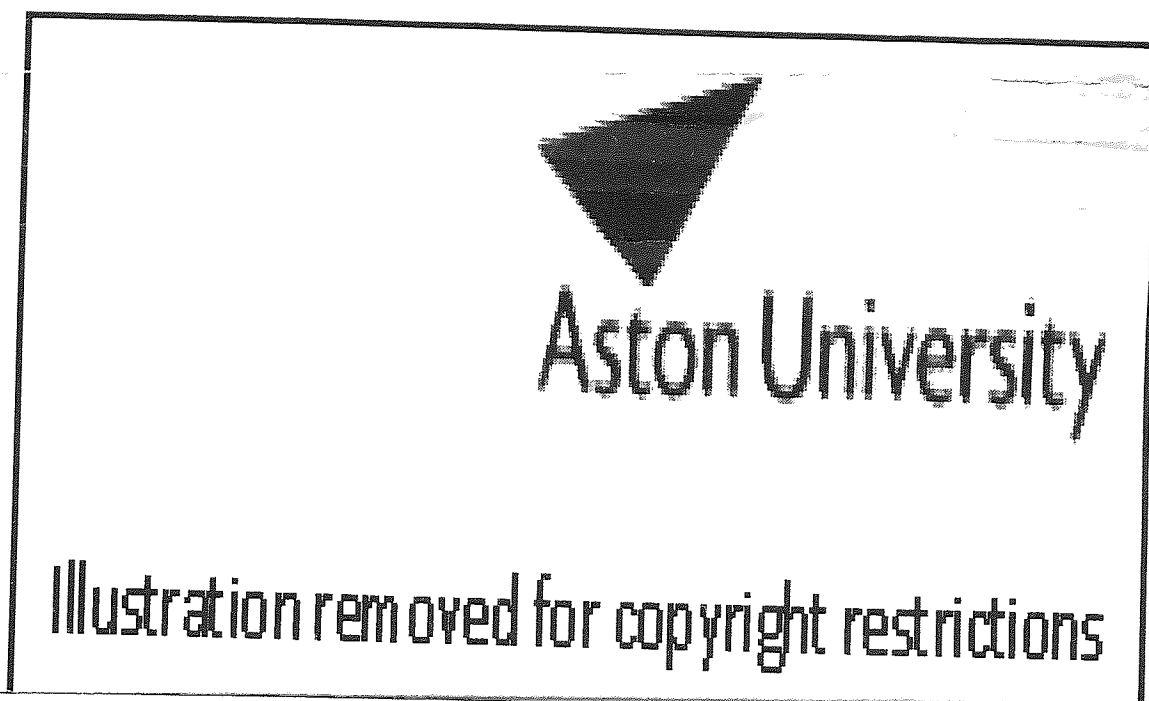


Figure 4.9: Implementation of Monaghan's refinement of Ladd's model (*Monaghan* (2003))

pitch values, and the normal width of the pitch range) are controlled by speaker-specific parameters. Others (e.g. the size of downstep, or the position of a tonal target within the pitch range) are controlled by language-specific parameter.

However, it has been shown that the model is quite unrevealing in some cases (*Ladd*, 1987). In other cases, large differences in the model parameters do not lead to striking audible difference in output speech. This limitation makes it difficult to use the model to relate intonational phenomena with perceptual consequence. This relation is crucial to accurate prosody modelling in the context of SY text-to-speech synthesis.

### 4.3 Summary

We have given a review of the most important intonation modelling techniques and paradigm in the context of tone language TTS. Based on the specification of our prosody model for SY TTS, none of the above intonation models is sufficient for our purpose. The data-driven models are inadequate due to their inflexibility and the inability to interpret their results linguistically. The Fujisaki model is particularly attractive, but

the phonological level of SY intonation cannot be easily modelled. The phonological rule-based approach proposed in *Ladd* (1990) and the computational approach proposed by *Gibbon* (2003a) are potentially capable of being used as the basis of a formal intonation system for SY prosody. However, the level of granularity we require between the phonological and phonetic level of speech prosody cannot be directly implemented within these models.

In the light of the above analysis, we consider a hybrid approach to be desirable. Noting the ability of Ladd's model in synthesising SY intonation and the computational strength underlying Gibbon's model, a way forward would be to integrate and refine these models, by: (i) incorporating perceptual measures into the model parameters, and (ii) employing a computational model within which intonation and prosody can be realised at an acceptable level of granularity.

# Chapter 5

## A review of duration models

The general goal of duration modelling in the context of TTS is to find a computational relation between a set of affecting factors and the time span allocated to the units of an utterance. The unit may be a word, syllable or segment of a syllable such as onset and rhyme. Duration modelling is a difficult task, primarily because of the complex relationships among interacting variables in natural speech. In models reported in the literature, specific variables that are assumed to influence the timing of speech such as the position of a spoken item in an utterance, etc., are usually identified. Natural speech is then analysed in the time domain, based on the identified variables, to determine the contribution of the variables to the perceptual quality of natural speech. However, it is impossible to totally isolate each speech variable due to the complex confounding nature exhibited by the component variables.

For example, the speaker characteristics or the influence of phonetic components that constitute a syllable are obscured by the contribution of syllable position in words, phrase and sentence. In order to determine priorities for the improvement of timing in synthetic speech, *Brinckmann and Trouvain* (2003) looked at the role of segmental duration prediction and the role of phonological symbolic representation in the perceptual quality of a text-to-speech system. *Córdoba et al.* (2002) used a neural network based approach to select the most significant parameters for duration modelling in Spanish TTS. The results of these studies suggest that the problem is a complicated one even when the acoustic unit and duration are determined based on the speech from a single person (*Plumpe and Meredith*, 1998). The complications arise because it is difficult to

make a speaker to produce speech sound with consistent acoustic behaviour from one utterance to another. Moreover, it has been shown that data measured from prepared laboratory speech will not necessarily be representative of natural speech, and a model that fits the duration of speaker **A** will not necessarily fit those of speaker **B** (*Campbell, 2000*).

Reported works in the literature (e.g. *O’Shaughnessy and Allen (1983)*; *Allen (1994)*; *Bellegarda et al. (2001)*; *Huckvale (2002)*) all agree that the present knowledge about speech duration, and the state of the art in speech technology in general, is still rudimentary and that our understanding of duration patterns and the many sources of variability which affect them is still sparse (*Möbius, 2003*).

There are four principal classes of methods applied in the design of duration models in the context of TTS systems (*Campbell, 2000*), they include: (i) look-up-tables, (ii) mathematical equations, (iii) rule-based, and (iv) data-driven. These duration models differ in structure as well as in the manner in which they use factors affecting duration to assign duration to units of utterance. The look-up-table method is used in small applications and also to augment the other methods by storing important duration data or rules for modelling rare events. The last three methods are more commonly used in modern TTS systems. In the following subsections, we review the works on these duration modelling methods in the context of TTS synthesis applications.

## 5.1 Rule-based duration model

In rule-based methods, (e.g. *Höhne et al. (1983)*) a set of If-Then rules are designed based on the durational pattern observed in a study of natural speech waveform. These rules are used to modify the duration of segments with the aim to produce a quality of match between natural and synthetic speech. Underpinning this approach is the idea that, by experimenting with a number of sentences and speakers, one could, hopefully, make a major improvement in the duration rules of the synthesisers and hence the quality of synthetic speech obtained.

Important literature in rule-based approaches to duration modelling are well documented in *Campbell (2000)*. For example, *Witten (1977)* described a higher-level

duration prediction algorithm for accommodating segment durations into predetermined foot timing. In the model, vowel onset locations are determined by the foot duration, subject only to modification for extreme cases of very long or very short foot. His rules were based on observations from speech but fails to take account of other factors known to influence the segment durations.

*Kohler* (1986) proposed a top-down timing model based on foot durations for German. His cascaded model computes syllable duration as a linear function of a predetermined foot duration, and assumes that segment duration are linear function of the syllable duration. The model for foot duration of the form:

$$R_{ijkl} = D + C_i + N_j + I_k + T_l \quad (5.1)$$

where  $R_{ijkl}$  is the rhythmic foot duration,  $D$  is a duration constant for a mono-syllabic foot at a neutral pitch pattern and a medium tempo,  $C_i$  is an additive duration constant for complexity in vowels and consonants in mono-syllabic foot.  $N_j$  is an additive constant for feet with more than one syllable.  $I_k$  is an additive constant for intonations other than neutral.  $T_j$  is an additive constant to allow for other speech rates. *Kohler's* model does not incorporate important duration phenomena such as final lengthening. This limits the applicability of the model for prosody modelling intended for TTS applications.

The *Klatt* (1987) duration model is perhaps the most popular rule-based duration model. This model predicts segmental duration by starting from some intrinsic value. The intrinsic duration is modified by successively applying rules which are intended to reflect contextual factors, such as positional and prosodic factors, that acts to lengthen or shorten the segment. The *Klatt* model assumes that: (i) each phonetic segment type has an inherent duration that is specified as one of its distinctive properties, (ii) it is possible to compute a percentage increase or decrease in the duration of the segment and (iii) segments cannot be compressed shorter than a certain minimum duration. The model is specified by the equation:

$$DUR = MINDUR + \frac{(INH DUR - MINDUR) \times PRCNT}{100.00} \quad (5.2)$$

where  $INH DUR$  and  $MINDUR$ , measured in milliseconds, are the inherent and



minimum duration of a segment respectively. *PRCNT* is the percentage shortening determined by applying rules.

A major weakness of this model is that rule parameters are determined by a manual trial-and-error process. Manually exploring the effect of mutual interactions among linguistic features of different levels is complex and highly prone to errors. Moreover, the model does not provide a systematic structure for determining how to include or exclude a duration affecting factor. Hence, the rule inferencing process usually involves a controlled experiment, in which only limited numbers of contextual factors are examined. In addition, the application of this model in a syllable based duration model such as ours will require that we treat syllables as segments. This will increase the complexity of our model because syllable-sized durations are generally less variable than sub-syllabic duration (*Keller and Zellner, 1995*). Using this approach will also introduce some inaccuracies in the representation of  $f_0$  anchor points which are crucial to the location of  $f_0$  peaks and valleys on the intonation contour of our prosody model.

## 5.2 Mathematical equation-based duration models

Mathematical equation based methods apply additive and multiplicative operations in the computation of speech unit duration. The sum-of-product (SOP) model (*van Santen, 1994; van Santen, 1992*) is a classical example which has been used in many TTS applications. The idea underlying the design of this model is that the regularity in the interaction of factors affecting duration can be described by a class of simple arithmetic models. An SOP model treats the factors as independent variables in a formula that computes a dependent variable, i.e. duration.

For example, assume that there are  $N$  factors affecting duration. Let these factors be denoted by the scale vector  $S_i$  which quantifies the duration effects associated with the  $i^{th}$  factors  $F_i, 1 \leq i \leq N$ . For example, if the factor  $F_1$  corresponds to the tonal features of SY syllables, it might comprise of three levels, “High”, “Mid” and “Low”. In the simplest case, the effect of this factor on duration can be captured by a vector comprising three elements, say  $S_1 = [S_{1,High}, S_{1,Mid}, S_{1,Low}]$ , with the appropriate values of  $S_{1,High} = S_1(High)$ ,  $S_{1,Mid} = S_1(Mid)$ , and  $S_{1,Low} = S_1(Low)$  estimated

from the data. If a given syllable is characterised by an input vector  $[f_1, f_2, \dots, f_N]$ , where each  $f_i$  represents the observed level of the associated factor  $F_i$ ,  $i \leq i \leq N$ , the duration  $D$  of a syllable can be formulated as follows (*Bellegarda et al.*, 2001):

$$F(D(f_1, f_2, \dots, f_N)) = G(S_1(f_1), S_2(f_2), \dots, S_N(F_N)) \quad (5.3)$$

where  $F(\cdot)$  and  $G(\cdot)$  are two arbitrary complex functions of the various duration involved. The SOP model is an equation that combines factor scales exclusively by forming sum and product using the formula:

$$DUR(\mathbf{f}) = \sum_{i \in T} \prod_{j \in I_i} S_{i,j}(f_j) \quad (5.4)$$

where  $T$  is some collection of indices associated with a subset of the set of factor,  $I_i$  is a collection of indices of factors occurring in the  $i^{th}$  subset and the scale function,  $S_{i,j}$ , maps from discrete to numerical values.

The SOP model assumes that  $F$  is a monotonically increasing transform, and that  $G$  can be decomposed as a sum-of-product of single parameters. The strength of the SOP model lies in the fact that the number of parameters required in the model are small, and that the arithmetic operation of multiplication and addition are sufficiently well behaved mathematically that parameters can be estimated even under conditions of severe frequency imbalance (*van Santen*, 1994; *Chung and Huckvale*, 2001).

However, *Bellegarda et al.* (2001) have observed that the diagnosis of an  $N$ -variable function on the basis of joint independence requires the testing of  $(N - 1)$ -tuples of variables for the independence of the  $N^{th}$ . Such diagnoses is not always successful because, apart from the fact that it will require a considerable effort in generating the model, it has been shown that  $D(F_1, f_2, \dots, f_N)$  is not a generalised additive function and the choice of the usual *log* function for  $F$  is probably not optimal (*Bellegarda et al.*, 2001).

### 5.3 Data-driven duration models

In the data-driven approach to duration modelling, the aim is to automatically generate a duration model from a large annotated speech corpus, usually with the aid

of statistical methods such as Classification And Regression Tree (CART) (*Lee and Oh, 1999; Chung, 2002*), automatic machine learning technique such as Artificial Neural Networks (ANN) (*Fletcher and McVeigh, 1993; Chen et al., 1998; Vainio, 2001*), Bayesian (*Goubanova and Taylor, 2000*), or Hidden Markov Models (HMM) (*Levinson, 1986; Donovan, 1996*). The philosophy underlying the application of this approach is that by carefully hand labelling a set of test sentences, it is possible, in theory, to automatically obtain the parameters for a good model of duration (*Batůšek, 2002*). Data-driven approaches are able to capture many of the variances observed in the duration of segments, within a sufficiently complex model, without the need for a clear understanding of how or what factors are affecting the quantal unit (*Bellegarda et al., 2001*).

The first step in implementing data-driven duration models is to identify the structure of the model. This structure describes a kind of relationship between syllable duration and some factors affecting duration. The next step is to design a model based on the identified structure. Next, relevant data are carefully collected and annotated, and the designed model is trained on the collected data. The goal of the training is to automatically induce phonological rules from the data by implicitly memorising them in the model's parameters (in the case of CART) or weights (in the case of ANN and HMM).

*Goubanova and Taylor (2000)* presented a Bayesian-based probabilistic approach to modelling segmental duration in the context of text-to-speech systems. The model was proposed mainly to tackle data sparsity and factor interaction problems in segmental duration modelling. Segmental duration was modelled as a hybrid Bayesian Network (BN) consisting of discrete and continuous nodes: each node in the network represents a linguistic factor that affects segmental duration. It was argued that the approach was ideal for duration modelling because the basic topology of the model is flexible, which allows the model designer to use knowledge to determine the factors that can be considered independent.

The consequence of this is that factor interactions can be captured by indicating the causal relationships of the factors in the connectivity of the nodes in a directed acyclic graph. A quantitative analysis of the model showed that it produced a good result in

terms of RMSE and correlation values (*Goubanova and Taylor, 2000*). It was claimed that the result is better or comparable to those produced by SOP and CART models. For example, the application of the model in vowel duration produced the median RMSE value of 5 *ms* and a correlation of 0.94. A corresponding SOP model produced RMSE of 9 *ms* and a correlation value of 0.9; while corresponding CART produced RMSE of 20 *ms* and a correlation value of 0.70. The higher RMSE suggests that the model does not approximate the data accurately enough while the lower correlation suggests that there is a low linear relationship between the model and the actual data.

However, it is not clear how the data sparsity problem is solved by the model since the data can only represent finite events from a Large Number of Rare Events (LNRE) (*Möbius, 2003*). In fact, *Goubanova (2002)* found that the BN model is quite sensitive to the amount of data used for prediction, producing lower correlation values for smaller sets. *Goubanova (2002)* also stated that the consonant analysis used in the model is not an optimal solution for the consonant duration prediction.

CART is perhaps the most popular data-driven method for duration modelling in TTS application. The strength of CART lies in the fact that standard tools for their generation are widely available, and that the computed regression tree is interpretable, in contrast to other data-driven models. An additional strength of CART is the ease with which trees may be built from duration data and from the speed of classification of new data. CARTs have been shown to cope with complex confounding interaction between factors affecting duration. This is because it makes very few assumptions about the structure of the data.

A CART embodies a binary tree with questions about the influencing factors at the nodes and predicted values at the leaves (*Breiman et al., 1984; Riley, 1992*). The tree itself contains *yes/no* questions about features and ultimately provides either a probability distribution, when predicting categorical values (classification tree), or a mean and standard deviation when predicting continuous values (regression tree). Well-defined techniques can be used to construct an optimal tree from a set of training data. Furthermore, CART induced trees can easily be converted into rules by viewing all the nodes that lead from the root to a leaf as the antecedent of a rule and the corresponding leaf as the consequence.

The weakness of CART, however, lies in its inability to accurately extrapolate from known to unknown contexts. Also, CART allows either a single feature or a linear combination of features at each internal node. This makes CART, like other binary decision-tree algorithms, to be biased towards generality. This feature is well suited for large disjunct data, but not for small disjunct data as obtained in duration modelling. This limitation results in the well known fragmentation problem, whereby the set of examples belonging to a tree node gets smaller and smaller as the depth of the tree is increased, making it difficult to induce reliable rules from deeper levels of the tree. This problem is made worse by the sparsity of duration data, hence making the development of a robust and reliable model very difficult. In addition, CART is computationally expensive as it requires the generation of multiple auxiliary trees (*Safavian and Landgrebe, 1991*).

## 5.4 Other duration models

Hybrid methods that combine two or more of the above data-driven methods have been proposed. For example *Fackrell et al. (1999)* combined CART and ANN methods by cascading regression tree with ANN. A major shortcoming of such a method is in the complexity of their design as well as the difficulty involved in their evaluation and improvement. It is difficult to interpret the resulting model even when the system is apparently working well.

Other traditional regression methods such as linear regression and tree regression with categorical predictor variables have been proposed (*Bartkova and Sorin, 1987; Riley, 1992*). Although linear regression models can be constructed to bring about a reliable inference of the duration model with small amount of training data, it is unable to accurately capture the dependencies among factors affecting duration (*van Santen and Olive, 1990*). To solve this problem, *Iwahashi and Sagisaka (2000)* proposed a Constraint Tree Regression (CTR) model which includes both linear and tree regression as special cases, and extends them to efficiently represent both dependent and independent parameters.

In summary, the SOP and CART-based duration modelling techniques have emer-

ged as the most popular in TTS application. The SOP method has, however, been shown to be superior to CART-based method for a number of reasons (*van Santen*, 1994). First, it needs far fewer training data to reach asymptotic performance. Second, the asymptotic performance is better than that of CART. Third, the difference in performance grows with the discrepancy between training and test data. Finally, adding more training data does not necessarily improve the performance of CART-based duration models. The SOP model has been applied to model duration for many languages including American English (*van Santen*, 1994), German (*Möbius and van Santen*, 1996), Mandarin Chinese (*Shih and Ao*, 1997) and Japanese (*Venditti and van Santen*, 1998). Building a successful SOP-based duration model, however, requires large annotated speech corpora, sophisticated statistical tools, and the type of linguistic and phonetic knowledge that is incorporated in traditional sequential rule systems. Factor interaction also prevents simple additive regression models (*Kaiki et al.*, 1990; *Bellegarda et al.*, 2001), which have good extrapolation properties, from being an effective solution.

## 5.5 Summary

In this Chapter, we have summarised the important literature on duration modelling in the context of TTS application, namely: rule-based, mathematical equation-based and data-driven approaches. As stated above, the main advantage that CRT and CART have over the other data-driven methods, such as neural networks and linear regression, is that their output is more readable and often understandable by humans. On the other hand, SOP models have been shown to successfully handle the data sparsity problem, which is a major weakness of CART. Therefore, a model that represents a half-way between CART, which is probabilistic and very assumptions, and SOP models, which are very prescriptive, is desirable. A Fuzzy Decision Tree (FDT) has the potential to adequately model this problem. We therefore propose to use the FDT technique for our duration modelling.

## Part III

# Design Methodology

# Chapter 6

## Design tools and techniques

In order to realise the prosody model proposed in this thesis, the tools and techniques for representing and manipulation knowledge in speech communication domain are required. In this chapter we highlight the tools and techniques used in our prosody model. They include: the relational tree technique, fuzzy logic and fuzzy control, classification and regression tree, fuzzy decision trees and text markup system.

### 6.1 The relational tree

The concept of representing a waveform in the form of a tree was first proposed by *Ehrich and Foith (1976)*. Underlying this concept is the idea that a waveform can be represented by its peaks and valleys and the relation between them. This representation can be modelled using a tree data structure in which the nodes of the tree correspond to the peaks and valleys in the waveform and the structure of the tree models the spatial structure of the waveform. This model is based on the assumption that, for most purposes, the necessary information about a waveform can be found in the peaks and the valleys of the waveform and the relationship amongst them. Representing a waveform by a tree is an intuitive process: a peak of a waveform is itself a waveform that contains a series of smaller peaks and valleys. And by analogy, a node on a tree is itself a root of a subtree. The important information contained in the tree can be easily accessed, extracted and modified in order to manipulate the waveform. The tree representation algorithm is generally deterministic so there is no need for backtracking.



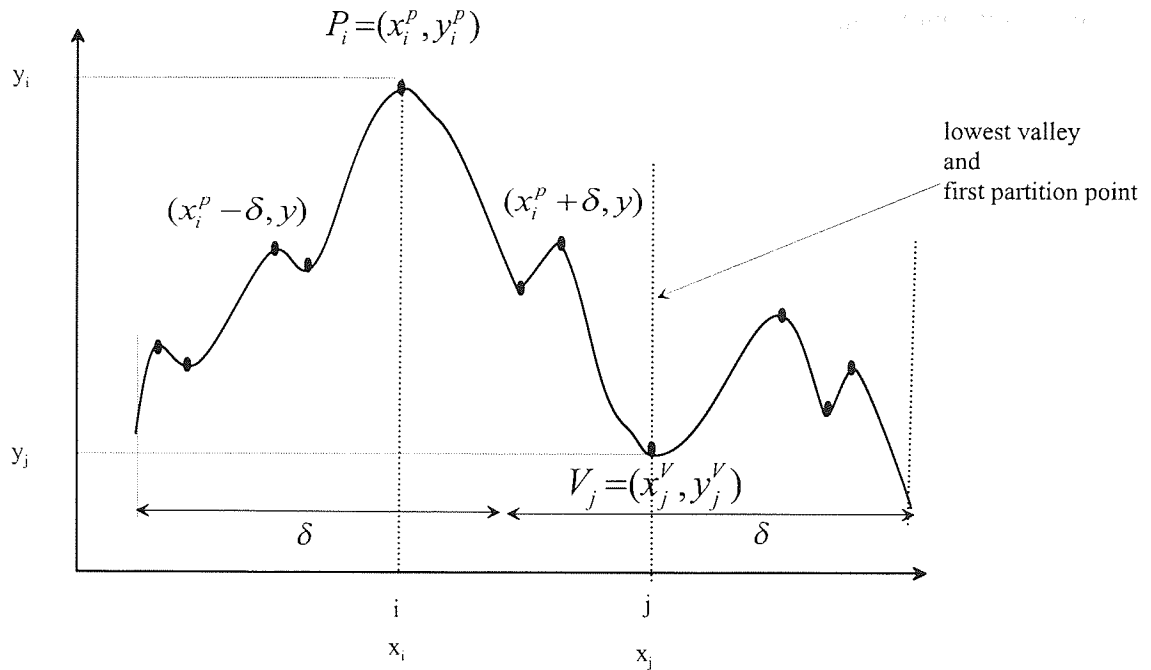


Figure 6.1: Peaks and valleys on a waveform

The peaks and valleys are relations that are defined over local windows on the waveform to be represented. Consider the waveform in Figure 6.1. The peak,  $P_i$ , and valley,  $V_j$ , occurs at point  $i$  and  $j$  respectively. A formal definition of these points over the waveform is as follows:

$$(x_i, y_i) \text{ is a peak, } P_i, \text{ if and only if } y_i \geq y : \quad \forall(x, y) |x_i - \delta| \leq x \leq |x_i + \delta| \quad (6.1)$$

$$(x_j, y_j) \text{ is a valley, } V_j, \text{ if and only if } y_j \leq y : \quad \forall(x, y) |x_j - \delta| \leq x \leq |x_j + \delta| \quad (6.2)$$

where  $2 \times \delta$  is the length of the window on the  $x$  axis. The value selected for  $\delta$  depends on the frequency of the waveform to be represented. If  $\delta$  is too large, some peaks or valleys will be missed. If, on the other hand,  $\delta$  is too small, the representation will contain a lot of redundant points.

If the waveform is scanned from left to right, a sequence of peaks and valleys will be encountered. Every peak,  $P_i$ , has definite left and right boundaries determined by the minima on each side, i.e. its right minimum  $V_i$  and its left minimum  $V_{1-i}$ . Similarly,

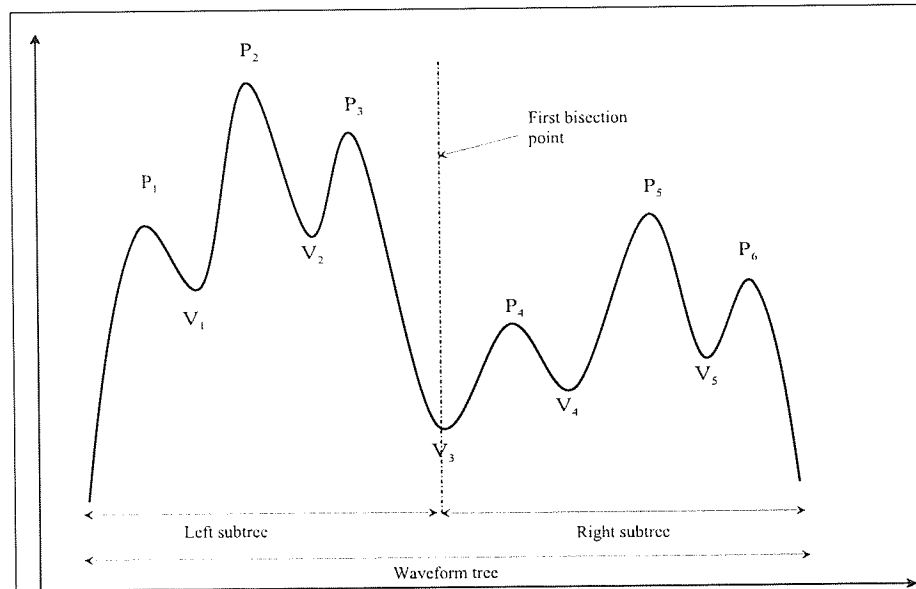
each valley,  $V_j$ , segments the waveform into two sub-waveforms which may contain a left peak,  $P_{j-1}$ , and a right peak,  $P_j$ . These peaks may be structured into several peaks by other valleys higher than the first.

*Cheng and Lu* (1985) proposed the following algorithm for constructing a tree representation of a waveform. First, a tree node is created and designated as the root of the tree. This node is labelled with the highest peak on the waveform. The waveform is partitioned into two sub-waveforms at the bottom of the lowest valley. For each sub-waveform, a node is created and assigned to be a child node of the root. This process is applied repeatedly to each sub-waveform until no more valleys can be found. The nodes generated at the end of the recursion are the leaf nodes of the tree and are labelled with the corresponding peaks.

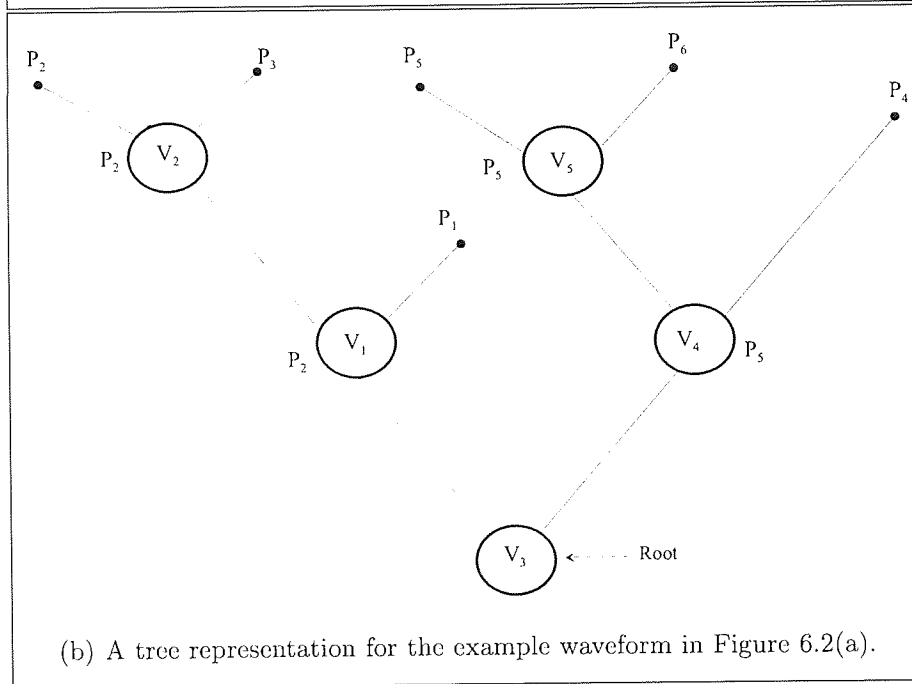
Figure 6.2(b) is a tree representation of the example waveform depicted in Figure 6.2(a). The deepest valley on the waveform shown in Figure 6.2(a) is  $v_3$ . This valley divides the waveform into two sub-waveforms. If we partition the waveform at this valley, we will have the left and right sub-waveforms indicated in Figure 6.2(a). The left sub-waveform contains peaks  $P_1$ ,  $P_2$ , and  $P_3$  while the right sub-waveform contains peaks  $P_4$ ,  $P_5$  and  $P_6$ . Peaks  $P_2$  and  $P_5$  are the dominant (highest) peaks on the left and right sub-waveform respectively.

The algorithm described above represents the abstract structure of the waveform only. This is, therefore, the skeleton of the waveform or the skeletal tree (S-Tree). The tree in Figure 6.2(b) is the S-Tree representation of the waveform in Figure 6.2(a). The S-Tree representation of a waveform has the following properties. The terminal nodes, i.e. leaves, of the tree correspond to peaks that have no further substructure. A non-terminal node corresponds to a valley and is labelled with the dominant peak of the valley. A non-terminal node and all the terminal nodes rooted on it correspond to a portion of the waveform being represented. The S-Tree representation of a waveform is not unique because two dissimilar waveforms can be represented using the same S-Tree.

The *Relational Tree* (R-Tree) representation of a waveform is a graph whose topological structure reflects structures of nested peaks. Nesting of the peaks is induced by sequences of valleys with increasing heights whose extents are enclosed by the extent of the next lower valleys (*Ehrlich and Foith*, 1976) and represented numerically. To



(a) Example waveform



(b) A tree representation for the example waveform in Figure 6.2(a).

Figure 6.2: Illustration of S-Tree generation

generate a complete *Relational Tree* (R-Tree), therefore, quantitative information such as amplitude, duration and intensity must be included in the S-Tree by adding attributes to the tree nodes. To achieve this, the co-ordinate values corresponding to the peaks, i.e.  $\mathbf{P} = \{P_1, P_2, \dots, P_n\}$ , and valleys,  $\mathbf{V} = \{V_1, V_2, \dots, V_n\}$ , of the dimension of interest, are computed.

There are a number of techniques for achieving this computation. One of them is to use a look-up table (*Ehrich and Foith, 1976*), e.g. a hash-table, which relates position and the context of the symbols to numerical data. Another approach is to apply a mathematical equation that uses initial values and the parameters of the S-Tree to compute the numerical values of  $\mathbf{P}$  and  $\mathbf{V}$  for each point. *Cheng and Lu (1985)* proposed an approach based on quantisation technique for solving this problem. In the quantisation technique, a grid is drawn over the waveform so that each set of crossings (locations where the grid touches the waveform) enclosed by the waveform on a quantisation level delineates an interval.

For the first quantisation level, the interval is the entire waveform. A node is created for this interval and designated the root node. Tree nodes are created for the interval from the current node to the intervals on the next quantisation level. This process is repeated until no further crossing can be found on the grid. The size of the skeletal tree is proportional to the complexity of the waveform being represented. Suppose that a waveform consists of  $m$  peaks; then the upper bound of the size of its skeletal tree is  $l \times m$ , where  $l$  is the number of quantisation levels, or the number of unique linguistic events.

In a two dimensional waveform, each of the vectors  $\mathbf{P}$  and  $\mathbf{V}$  has exactly one dimension. In an  $n$  - dimensional waveform ( $n > 2$ ),  $\mathbf{P}$  is an  $n - 1$  dimensional vector while  $\mathbf{V}$  is a single dimensional vector. The variable  $\mathbf{V}$  may be considered as the independent variable, e.g. time, and  $\mathbf{P}$  is a vector of dependent variables. After the computation of the points, interpolation techniques can be applied to join the points in order to obtain a continuous waveform.

In the past, the R-Tree technology has been applied to the solution of various problems such as recursive or cascade logical filter, texture description, robotics (*Ehrich and Foith, 1976*), waveform matching (*Cheng and Lu, 1985*) and ECG classification

(Shaw and Defigueiredo, 1990). To the best of our knowledge, this is the first work to apply the R-Tree technology to intonation and prosody modelling.

## 6.2 Fuzzy logic and fuzzy control

A fuzzy set  $A$  of a universe of discourse  $U$  is characterised by a membership function  $\mu_A$ . This function is defined as:

$$\mu_A(x) : U \rightarrow [0, 1] \quad (6.3)$$

where each element  $x$  of  $U$  has a number  $\mu_A(x)$  in the interval  $[0.0, 1.0]$  which represents the degree to which  $x$  belongs to  $A$ . A  $\mu_A$  value of 1.0 indicates full membership and a value of 0.0 indicated non-membership. A  $\mu_A$  value of 0.5 indicates that  $x$  belongs to  $A$  at a degree corresponding to 50% membership.

The fuzzy set theory employs the membership function to interpret uncertain situation and imprecise information. In order to manipulate the fuzzy sets in a manner similar to ordinary sets that uses Boolean operations, Zadeh (1972) proposes the extension of ordinary set theory to fuzzy sets. The notion of complement (NOT) intersection (AND), and union (OR) operations for fuzzy sets are defined as follows:

**Complement (NOT):**

$$\mu_{\bar{A}}(x) = 1.0 - \mu_A(x). \quad (6.4)$$

**Intersection (AND):**

$$\mu_C(x) = \mu_A(x) \wedge \mu_B(x) = \min[\mu_A(x), \mu_B(x)]. \quad (6.5)$$

**Union (OR):**

$$\mu_C(x) = \mu_A(x) \vee \mu_B(x) = \max[\mu_A(x), \mu_B(x)]. \quad (6.6)$$

The use of fuzzy sets provides a powerful tool for extending the capability of binary logic in ways that enable a much better representation of linguistic knowledge. Fuzzy logic differs from probability in that probability deals with *randomness* of future events, while fuzzy logic uses *possibility* theory to deal with the impression of current or past event (Pal and Majumder, 1977).

Fuzzy control is the most common application of fuzzy logic in intelligent systems engineering. This is partly because fuzzy logic, on which fuzzy control is based, is much closer to human thinking and natural language than traditional logic systems, which involves computing with numbers and symbols which are crisp. Fuzzy logic provides an effective means for capturing the approximate, inexact nature of the real world (Lee, 1990a). The essential part of the fuzzy controller is a set of linguistic control rules related by dual concepts of fuzzy implication and the compositional rule of inference. A fuzzy model is viewed as a linguistic descriptor by virtue of fuzzy logic proposition. The description uses *linguistic variables* and takes advantage of a rule base of expert knowledge written in the form of linguistic expression.

A linguistic variable is characterised by a 5-tuple (Lee, 1990a):  $(x, T(x), U, G, M)$ , where  $x$  is the name of the variable;  $T(x)$  is the term set of  $x$ , i.e. the set of names of linguistic labels of  $x$  with each label being a fuzzy term defined on  $U$ , the universe of discourse.  $G$  is a syntactic rule for generating the names of values of  $x$ ; and  $M$  is a semantic rule for associating a meaning with each value.

For example, if we take *pitch* as a linguistic variable, then its term set,  $T(\text{pitch})$ , could be  $T(\text{pitch}) = \{\text{High}, \text{Mid}, \text{Low}\}$ , where each term in  $T(\text{pitch})$  is characterised by a fuzzy set in a universe of discourse,  $U = [75.0, 250.0]$ , over the fundamental frequency ( $f_0$ ) dimension. We might interpret the linguistic term “High” as “a pitch associated with an  $f_0$  peak above  $150.0\text{Hz}$ ”, “Low” as “a pitch associated with an  $f_0$  peak below  $80.0\text{Hz}$ ”, and “Mid” as “a pitch associated with an  $f_0$  between  $80.0\text{Hz}$  and  $150.0\text{Hz}$ ”. These terms can be characterised as fuzzy sets whose membership functions are shown in Figure 6.3.

Using this membership function, we can compute the degree to which any  $f_0$  value in the interval  $[75.00, 250.00]$  belongs to the pitch level High, Mid or Low. As illustrated in Figure 6.3, the  $f_0$  of  $76.00\text{Hz}$  belongs to a low pitch to a degree of 0.8 (i.e.  $\mu_{\text{Low}}(76.00) = 0.8$ ), 0.12 to Mid (i.e.  $\mu_{\text{Mid}}(76.00) = 0.12$ ) and has no membership in the High pitch range, (i.e.  $\mu_{\text{High}}(76.00) = 0.0$ ).

Fuzzy control logic can be used to construct an automatic control system to simulate the human *know-how* operation. The basic structure of a Fuzzy Logic Controller (FLC) is depicted in Figure 6.4. It comprises four basic elements: a *fuzzification interface*, a

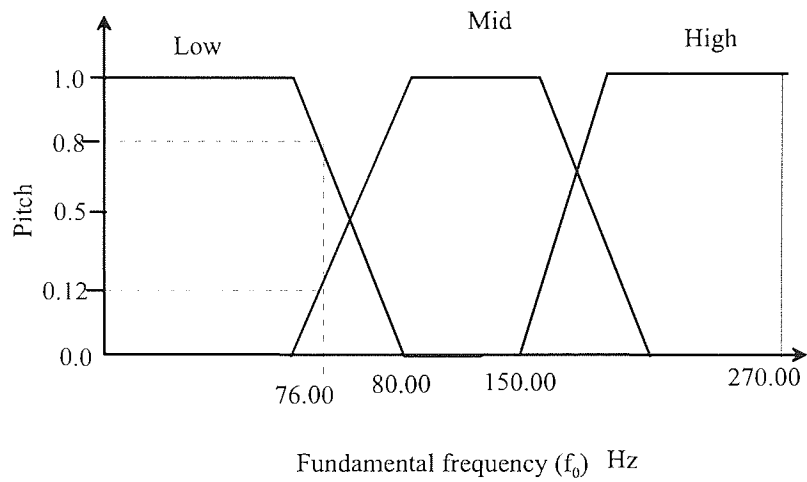


Figure 6.3: Example of fuzzy set

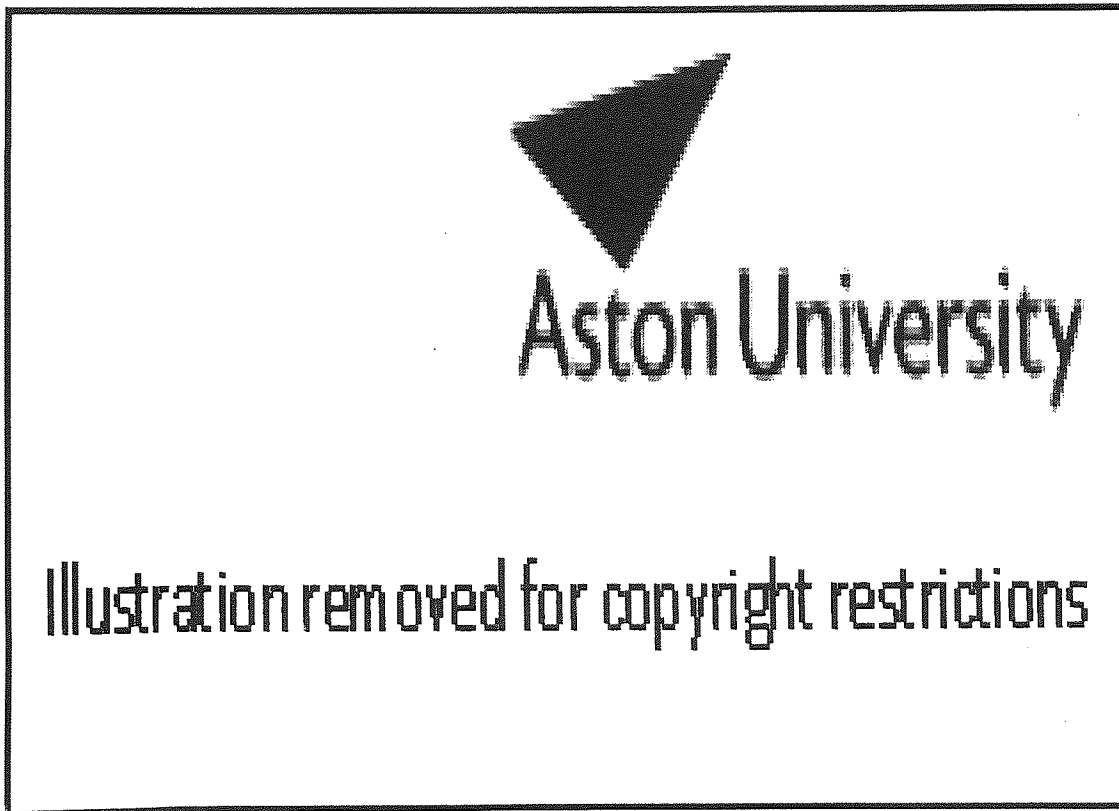


Figure 6.4: Fuzzy logic Controller (*Lee (1990a)*)

*knowledge base*, *decision logic*, and a *defuzzification interface*.

*fuzzification interface* performs the function that converts crisp input data into suitable linguistic values. This is achieved using the membership function. The universe of discourse is divided into several regions and each region has a linguistic term. Each term has a membership function which is used to compute the degree to which an input crisp value belongs to the term. There are many fuzzification functions used in FLC, namely e.g. triangular, trapezoidal, and Gaussian bell (Chen and Otto, 1995; Mitaim and Kosko, 2001).

*knowledge base* contain data that defines the linguistic control and rules that characterise the control strategies. The data provide necessary definitions which are used to define linguistic control rules and fuzzy data manipulation in a FLC. The rule base, on the other hand, determines the control goal and strategies as expressed by a human expert.

*decision logic* is the kernel of the FLC, its aim is to simulate human decision making based on fuzzy concepts and to infer fuzzy control actions by employing fuzzy implication and the rules of inference in fuzzy logic.

*defuzzification interface* performs a scale mapping from the linguistic domain to crisp values on the universe of discourse. This is the final stage in the fuzzy control system. There are a number of defuzzification methods including centre of area, maximum possibility, mean of maximum possibility, centre of mass of highest intersection region, etc. (Takagi and Sugeno, 1985; Castro and Delgado, 1996).

In a fuzzy logic controller, the dynamic behaviour of a system is characterised by a set of linguistic description rules based on expert knowledge. The expert knowledge is usually expressed in the form:

<p>IF (a set of conditions are satisfied) THEN (a set of consequences can be inferred).</p>
---



The antecedent and the consequence of the IF-THEN rules are associated with fuzzy linguistic terms. The fuzzy control model proposed by *Takagi and Sugeno (1985)* (TS) is used in our prosody modelling. This is a nonlinear model represented by the following equations:

$$\begin{aligned}
 R^i : & \text{ IF } ((x_1 \text{ IS } A_1^i) \text{ AND } (x_2 \text{ IS } A_2^i) \text{ AND } (\dots) \text{ AND } (x_m \text{ IS } A_m^i)) \\
 & \text{ THEN } y = p_0 + p_1x_1 + p_2x_2 + \dots + p_nx_n
 \end{aligned} \tag{6.7}$$

where  $R^i$  is the label for the  $i^{\text{th}}$  rule in the rule base. The implication,  $\hat{y}$ , of a set of  $m$  inputs, in a rule base comprising  $c$  rules is computed as:

$$\hat{y} = \frac{\sum_{i=1}^c w^i y^i}{\sum_{i=1}^c w^i} \tag{6.8}$$

where

$$w^i = \prod_{j=1}^m A_j^i(x_j) \tag{6.9}$$

Equation 6.7 uses linguistic relations expressed in the rule premise to predict numerical values expressed in the rule consequence. The model identification process for a TS-based model is divided into two sub-processes: (i) structure identification and (ii) parameter identification. In the structure identification process, the input (i.e.  $x_1, \dots, x_m$ ) and output (i.e.  $y$ ) variables of the model are identified based on the input-output data. The number of partitions, i.e. linguistic terms, required for the input variables (i.e.  $A_1, \dots, A_m$ ) are also determined. The model structure can also be identified based on a description of the system by an expert. In the parameter identification process, the numerical values of the consequent parameters, i.e.  $p_0, \dots, p_n$ , are determined.

The fuzzy logic based model is popularly used in complex problems where the input-output relations of the systems is ambiguous and uncertain, such as in intelligent process control. This includes fossil power plant modelling (*Arroyo-Figueroa et al., 2000*) chemical plant and gas furnace (*Sugeno and Yasukawa, 1993*) as well as many intelligent systems applications (*Lee, 1990b; Bezdek, 1993*).

Fuzzy logic has been applied in a number of studies in speech and language engineering. Examples include a number of works in speech recognition (*Pal and Majumder, 1977; De Mori et al., 1979; Demichelis et al., 1983; Mori et al., 1985; O'Brien, 1993; Yu and Oh, 1995*), speaker identification (*Castellano and Sridharan, 1996*), and speech signal processing (*Kosanović et al., 1996; Breining, 2001*). Results of these studies show that fuzzy logic based models are capable of modelling the acoustic parameters of speech signal.

In speech synthesis, *Raptis and Carayannis (1997)* demonstrated the application of a fuzzy logic rule based approach to formant speech synthesis. Also, *Jitca et al. (2002)* have shown that speech synthesis can be improved by applying a fuzzy logic based method to the computation of the first two formants (i.e.  $F1$  and  $F2$ ) of phonemes. Recently, *Lin et al. (2003)* combined a fuzzy logic based approach with a recurrent neural network for modelling Mandarin Chinese prosody.

In this work, we are using fuzzy logic based approach for duration and intonation modelling in the context of our relational tree based prosody model for SY speech synthesis.

### 6.3 Classification And Regression Tree (CART)

The concept of decision tree was popularised by *Quinlan (1986)* with the *Interactive Dichotomizer (ID3)* algorithm. Systems based on this approach use an information theoretical measure of entropy for assessing the discriminatory power of each attribute in a problem space. The main objectives of decision trees are to: (i) represent the training sample as correctly as possible, (ii) generalise beyond the training sample so that unseen (or test) samples could be predicted or classified with as high accuracy as the training sample, (iii) make the represented knowledge easily updatable as more training samples become available, (iv) have a simple structure that can be easily interpreted.

The main feature of decision trees is their capability to break down a complex decision-making process into a collection of simpler decisions, thereby providing an easily interpretable solution (*Mitra et al., 2002*).

Decision trees are attractive for solving complex problems for the following reasons (Quinlan, 1990; Safavian and Landgrebe, 1991):

1. Global complex decision regions can be approximated by the union of simpler local decision regions at various levels of the tree.
2. They eliminate unnecessary computation by testing training samples against only those subsets of classes that are most likely.
3. There is the flexibility of choosing different subsets of features at different internal nodes of the tree such that the feature subset chosen optimally discriminates among the classes in a node.
4. By using a smaller number of features at each internal node, decision trees are able to model multi-dimensional problems without excessive degradation in performance.
5. Prior partitioning is not required for continuous non-symbolic attributes.

Classification And Regression Tree (CART) (Breiman *et al.*, 1984; Riley, 1992) is the most popular decision tree used in prosody modelling in the context of TTS applications, particularly in duration modelling. The core idea of CART construction is to find an appropriate question, from a set of possible questions, about the input feature that makes the best split of data. The best question is usually determined using the *entropy* function which computes the information gain for that partition. A group containing more similar samples have lower entropy. After a split is made, the algorithm continues to make new partitions recursively by finding a question that maximises the entropy. The recursion stops when some criteria are met, such as the minimum number of samples in a partition.

Basically, the CART algorithm induces a decision tree from the data by recursively partitioning the data on the input variable that maximises the decrease in entropy of the training set. The classification performed by the CART algorithm is analogous to “data compression” of the training data set.

Given a training set consisting of  $N_{train}$  labelled examples  $\{(\mathbf{x}_n, y_n); n = 1, 2, \dots, N_{train}\}$ , where  $\mathbf{x}_n$  represents the independent and  $y_n$  represents the dependent variables. The training set is used to build a tree (T), composed of a collection of nodes ( $t_i \in T; i = 0, 1, 2, \dots$ ) arranged in a hierarchical manner. CART uses a top-down approach to build a binary tree. The first node of a CART, T, is the root node, denoted by

$t_0$ . By definition, all examples are assigned to this node. The tree is constructed by a divide-and-conquer (Quinlan, 1990) strategy in which the attribute space is partitioned by a hierarchy of Boolean tests into a set of non-overlapping regions. Each of the tests in the hierarchy corresponds to an internal node of the decision tree and each node represents a simpler decision.

In the classification problem, the CART selects the split that leads to the largest decrease in the impurity or classification error of the tree. Other local optimisation criteria, such as information gain and information-gain ratio (Mittra *et al.*, 2002), may also be used. In a regression problem, the value of  $y$  which is the decision assigned by the tree to a given vector of attributes, i.e.  $\mathbf{x}_{test}$ , is given by the equation:

$$\bar{y}(\mathbf{x}_{test}) = \sum_{t_i \in \tilde{T}} \mu_i(\mathbf{x}_{test}) \bar{y}_i \quad (6.10)$$

where  $\bar{y}_i$  is given by:

$$\bar{y}_i = \frac{1.0}{N_i} \sum_{n=1}^{N_{train}} \mu_i(\mathbf{x}) y_n \quad (6.11)$$

$\mu_i(\mathbf{x})$  is a function which evaluates to 1 for those examples that satisfy the conjunction of Boolean tests leading to  $t_i$ , and 0 otherwise. The error rate of the tree on the training set is given by the equation:

$$R_{train}(T) = \frac{1.0}{N_{train}} \sum_{n=1}^{N_{train}} (y_n - \bar{y}(\mathbf{x}_n)) \quad (6.12)$$

A number of researchers have applied the CART technique in duration, intonation as well as prosody modelling. This include, in duration modelling: *Brinckmann and Trouvain* (2003) for German TTS, *Chung and Huckvale* (2001) and *Chung* (2002) for Korean TTS, *Batůšek* (2002) for Czech, and *Bouzon and Hirst* (2002) for British English; in fundamental frequency prediction, *Ljolje and Fallside* (1986); and in prosody modelling, *Lee and Oh* (1999) and *Viana et al.* (2003).

In *Chung and Huckvale* (2001), for example, the phonetic and phonological factors affecting the rhythm and timing in spoken Korean was modelled by a stepwise construction of CART using *Wagon*. They reported an RMSE of 24.23ms and a correlation of 0.77 on training set. On the test set, they reported an RMSE of 26.48ms and

a correlation of 0.73. *Lee and Oh* (1999) proposed a tree-based modelling of prosodic phrasing, pause duration between phrases and segmental duration for Korean TTS system and they reported a correlation of 0.82.

*Dusterhoff et al.* (1999) used CART based on the Tilt intonation modelling framework for modelling English intonation. For quantitative evaluation, they reported an RMSE=34.3 and a correlation of 0.6. *Viana et al.* (2003) proposed a prosody model for European Portuguese based on hand-annotated text and using CART. In this model, two questions were used to represent the part-of-speech (POS) information, namely: (i) the number of different tags in the utterance, (ii) the number of words that must be included in the analysis window. The distance measure in terms of words from the boundary to previous and following punctuation marks were also taken into account. The major decision factor was the location of the punctuation marks and the POS tags for the word. A CART was then trained using 41 POS tags.

All of the above mentioned studies show that CART is capable of modelling speech data reasonably well. However, CART has a number of limitations which renders it unsuitable for our purpose. First, the CART construction methodology uses a *greedy algorithm* whereby it finds the locally best questions when making a split. This approach is well known for its sub-optimal approximation (*Boyer and Wehenkel*, 1999). Second, CART allows only either a single feature or a linear combination of features at each internal node. This limits the degree of freedom that can be applied to the attribute being queried. Third, CART is computationally expensive as it requires the generation of multiple auxiliary trees. Fourth, CART selects the final pruned subtree from a parametric family of pruned subtrees, and this parametric family may not include the optimal pruned subtree (*Safavian and Landgrebe*, 1991). Last, and most importantly, the intervals involved in quantitative association as represented by CART is not concise and meaningful enough for human experts to obtain non-trivial knowledge which are expressed linguistically.

Knowledge expressed in linguistic representation is natural and much easier to comprehend by humans. Fuzzy decision trees have the capacity to represent this kind of knowledge.

## 6.4 Fuzzy Decision Tree (FDT)

Fuzzy Decision Tree (FDT) is another variant of the ID3 algorithm. It exploits fuzzy logic algorithms in the construction of a decision tree. The major strengths of fuzzy logic algorithms are that they are robust and flexible and they are able to cope well with the interactions of several linguistic attributes. Hence, they can be more easily tailored for coping with small disjuncts, which are associated with large degrees of attribute interaction (*Carvalho and Freitas*, 2002). When this attribute is coupled with the tree data structure, the result is a powerful and robust model. An attribute of such a model is its explicit structure resulting in the ease of interpretation of represented knowledge. The fuzzy decision tree (FDT) approach combines symbolic decision trees with the approximate reasoning offered by fuzzy logic.

FDT has been applied to the solution of many problems. In health informatics, this includes the modelling of diseases such as diabetes, cancer and heart failure (*Suárez and Lutsko*, 1999; *Olaru and Wehenkel*, 2003); in intelligent process control, FDT has been used in modelling a robot soccer system (*Sison and Chong*, 1994; *Huang and Liang*, 2002); in software engineering (*Pedrycz and Sosnowski*, 2001), in weather prediction *Dong and Kothari* (2001) as well as power system security assessment (*Boyen and Wehenkel*, 1999).

Although CART has been widely used in duration modelling, our literature review shows that FDT has not been applied to this problem. The present work, therefore, is the first to apply FDT in this context. Specifically, we are using the FDT in duration modelling in the context of our prosody modelling for TTS.

In the past, FDT were constructed by first constructing a CART and replacing its binary decision using fuzzy logic. For example, *Suárez and Lutsko* (1999) proposed a globally optimal FDT for classification and regression by starting from CART, whose binary split are replaced by fuzzy splits. The parameter of the fuzzy split are then determined by the global optimisation of cost function. By using the skeleton of the CART, *Suárez and Lutsko* (1999) took the advantage of well-studied heuristics developed to extract the tree architecture directly from the training data. In this way, they are able to avoid limiting *a priori* the final size, depth, or complexity of the FDT. The use of CART in the design of FDT is also discussed by *Jang* (1994).

Given an  $N$  labelled data set partitioned into  $l$  sets of patterns belonging to classes  $C_i$ ,  $i = 1, 2, 3, \dots, l$ . Let the population in class  $C_i$  be  $n_i$ . Each pattern has  $n$  features and each feature can take on two or more values. The ID3 algorithm for generating an efficient decision tree can be stated as follows (Quinlan, 1990; Mitra et al., 2002):

1. Calculate the initial value of entropy using the equation:

$$Entropy = \sum_{i=1}^l - \left( \frac{n_i}{N} \right) \log_2 \left( \frac{n_i}{N} \right) = \sum_{i=1}^l -p_i \log_2 p_i. \quad (6.13)$$

2. Select the feature which results in the maximum decrease in entropy (gain in information), to serve as the root node of the decision tree.
3. Build the next level of the decision tree by using the attribute that provides the greatest decrease in entropy.
4. Repeat steps 1 through 3 until all subpopulation are of a single class and the system entropy is zero.

In order to incorporate fuzziness into the tree, three steps are required: (i) the input attributes are discretised in linguistic terms. This is done by specifying fuzzy labels on the input domain and defining a fuzzy membership function for each of the terms, (ii) definition of a fuzzy entropy function that computes entropy at the node level, in terms of class membership. The fuzzy entropy considers the membership of a pattern to a class and helps enhance the discriminative power of an attribute. A typical fuzzy entropy function is as follows (Janikow, 1993):

$$Entropy = - \sum_{i=1}^l - \left( \frac{\sum_{j=1}^N \mu_{ij}}{N} \right) \log_2 \left( \frac{\sum_{j=1}^N \mu_{ij}}{N} \right) \quad (6.14)$$

and (iii) a defuzzification function for inferencing.

An FDT is similar to a CART but they differ in the following respects:

1. Data samples may match more than one test of a node in FDT. When aggregated over multiple level, this leads to samples falling into many nodes, with a real-value degree of membership.
2. The information content in FDT implements fuzzy partial memberships rather than the binary division in CART.
3. Fuzzy match is determined based on pre-selected norms.

The fusion of fuzzy sets with decision trees enables one to combine the uncertainty handling and approximate reasoning capabilities of the former with the comprehensibility and ease of application of the later. This enhances the representation power of decision trees with the knowledge component inherent in fuzzy logic. This leads to a better robustness, noise immunity, and applicability in uncertain/imprecise domain such as prosody modelling.

Fuzzy decision trees assume that all domain attributes or linguistic variables have pre-defined fuzzy terms which impose a fuzzy restriction on their values. The information gain measure function is modified for fuzzy representation and a pattern can have non-zero match to one or more leaves.

## 6.5 Text markup in TTS

When a piece of text is read aloud, how much information from the text is responsible for the sound-waveform generated? The text certainly conveys clues to the grouping of syllables in word sequences. It also contains cues about important syntactic boundaries that can be identified by way of punctuation marks. SY text, if written with the complete orthographic specification, contains cues about tones and syllables as well as their organisation into words, phrases and sentences. The identification and appropriate use of these cues can greatly improve the computation of an accurate intonation for the text.

In general, however, such syntactic cues are not enough to facilitate the generation of adequate prosody from text. For example, the same sequence of words can be spoken in different ways, depending on the context. What really matters is the marking of structure and contents in such a way that pauses and emphases are placed correctly and the hierarchy of phrasing and prominence is equitably conveyed (*Monaghan, 2001*). *Taylor and Isard (1997)* have observed that plain text is the most desirable form of input to a TTS system from a human perspective due to its standard nature and universal understanding. Therefore, there is the need to render input text in such a way that it can be easily read by a human as well as easily processed by a TTS system.

The best way to achieve this goal is to explicitly annotate the input text with



information that will aid further processing of the text. The idea of a text markup language was first introduced by *Goldfarb et al.* (1970) with the design of the Generalised Markup Language (GML), which later evolved into the Standard Generalised Markup Language (SGML). Since then, a number of text markup languages have been developed and used.

Many TTS developers have designed text Markup Languages (ML) specifically for their TTS applications (e.g. *Taylor and Isard* (1997); *Huckvale* (1999, 2001)). Some of these ML include: Spoken Text Markup Language (STML) and Speech Synthesis Markup Language (SSML) (*Taylor and Isard*, 1997; *Burnett et al.*, 2002), VoiceXML (*VoiceXML*, 2000) as well as the Java Speech Markup Language (JSML) (*JavaSpeechML*, 1997). JSML was developed to facilitate a standard text markup and programming for TTS engine in the Java environment. Most of these mark-up languages provide text description tags that describe the structure of the document, and speaker directive tags that control the emphasis, pitch rate, and pronunciation of the text.

SABLE (*Sproat et al.*, 1998) is a TTS markup language developed by combining STML and two other markup languages, i.e. Java Speech Markup Language (JSML) and Speech Synthesis Markup Language (SSML), to form a single standard. It is designed to be easily implemented in any TTS engine. It has been implemented in both the Edinburgh University's Festival speech synthesis system and the Bell Labs TTS engine (*Taylor and Isard*, 1997).

The following is a simple example of SABLE markup for SY two sentence paragraph: ““Bàbá àgbè ti ta cocoa 30 kg ní N500 kí ó tó mò pé àjò NCB ti fi owólé cocoa. Nì ò'kan bíi dédédé agogo 3:00 ọsán ni bàbá àgbè délé” (meaning “[Father farmer has sold cocoa 30 kg before he knows that organisation NCB has add money to cocoa. At about time 3:00 afternoon father farmer got home] The farmer has sold 30 kg of cocoa for N500 before realising that the NCB organisation has increased cocoa price. The farmer got home around 3:00 in the afternoon”):

```
<DIV TYPE="paragraph">
  <DIV TYPE="sentence" >
    Baba agbe ti ta cocoa 30 kg ni N500 ki o to mo pe ajo
    NCB ti fi owole cocoa.
  </DIV>
</DIV TYPE="sentence">
```

Ni n'kan bii dede agogo 3:00 osan ni baba agbe dele.

</DIV>

</DIV>

In this example, the structure of the text to be pronounced is clearly marked. SABLE also includes tags to specify numeral and other text anomalies.

The markup systems discussed above are not suitable for SY text because they do not adequately describe data and structures found in typical SY text. Using them for marking SY text will lead to a complex representation system which is difficult to use. An alternative suggested by *Huckvale* (2001) is to provide an open, non-propriety textual representation of the data structures at every level and stage of processing. In this way, additional or alternative components may be easily added even if they are encoded in different format. This approach was used in the *ProSynth* project (*Ogden et al.*, 2000).

In *ProSynth*, each composed utterance comprising a single intonation phrase is stored in a hierarchy. Syllables are cross-linked to the word nodes using linking attributes. This allows for phonetic interpretation rules to be sensitive to grammatical function of a word as well as to the position of the syllable in the word. Knowledge for phonetic interpretation is expressed in a declarative form that operates on prosodic structures. A special language called *ProXML* was used to represent knowledge which is expressed as unordered rules and it operates solely by manipulating the attribute on XML-encoded phonological structure.

The *ProXML* implementation suggests that the facility provided by XML matches the requirements to represent the phonological features of an utterance in a metrical prosodic structure, namely: nodes described by attribute-value pairs forming strict hierarchies (*Huckvale*, 1999). An important attribute of this structure for prosody modelling is that the phonetic descriptions and timings can be used to select speech unit and expresses their durations and pitch contour for output with a TTS system.

The apparent success of XML at representing phonological knowledge, as well as the additional advantage that the represented text can be published on the Internet, motivated our use of XML in developing the markup system for the SY language. The detailed design of our XML based SY text markup system is presented in Appendix C.

## 6.6 Summary

In this chapter, we have discussed the fundamental technologies and techniques central to the development of our prosody model for the Standard Yorùbá language. The techniques discussed include the *Relational tree* technique which will be used to model the fundamental structure of our prosody model. Underlying the technique is the idea of representing a waveform in the form of a tree. Also discussed is the fuzzy logic technique which will be used to compute the acoustic parameter on the fundamental frequency dimension of our prosody model. Next, we discussed Classification and Regression Tree (CART) and Fuzzy Decision Tree (FDT), these will be used to model the duration dimension. We also reviewed literature on text markup systems and discussed the reasons underlying our selection of XML for implementing our text markup system.

# Chapter 7

## Model conceptualisation and design

### 7.1 Overview of the SY TTS system

The general structure of the proposed SY TTS systems is shown in Figure 7.1. The SY text to be input is first typeset using  $\text{\LaTeX}$ . The XML tags are then markup over the  $\text{\LaTeX}$  text. The function of the markup process is to annotate important prosodic cues. The tags also provide information which facilitates the unambiguous identification of prosodic events in the input text, e.g. phrase and sentence boundaries. The tags provide information that is used to compute the abstract structure of intonation and subsequent computation of the various dimension of prosody. The synthetic SY speech is generated from the annotated text in a number of modular steps. These steps are briefly discussed as follows:

**Text analysis module:** The text analysis process segments the marked up text input and handles the ambiguities associated with the use of punctuation marks and related symbols. Its specific function is to identify textual anomalies such as numerals, abbreviations and acronyms that must be expanded into their lexical equivalents by subsequent processing steps.

**Text normalisation module:** The text normalisation process expands all textual anomalies in the text using the information provided by the tags in the markup as well as the identification of the input as provided by the text analysis process. At the end of this process, the text will contain only letters of the SY alphabet

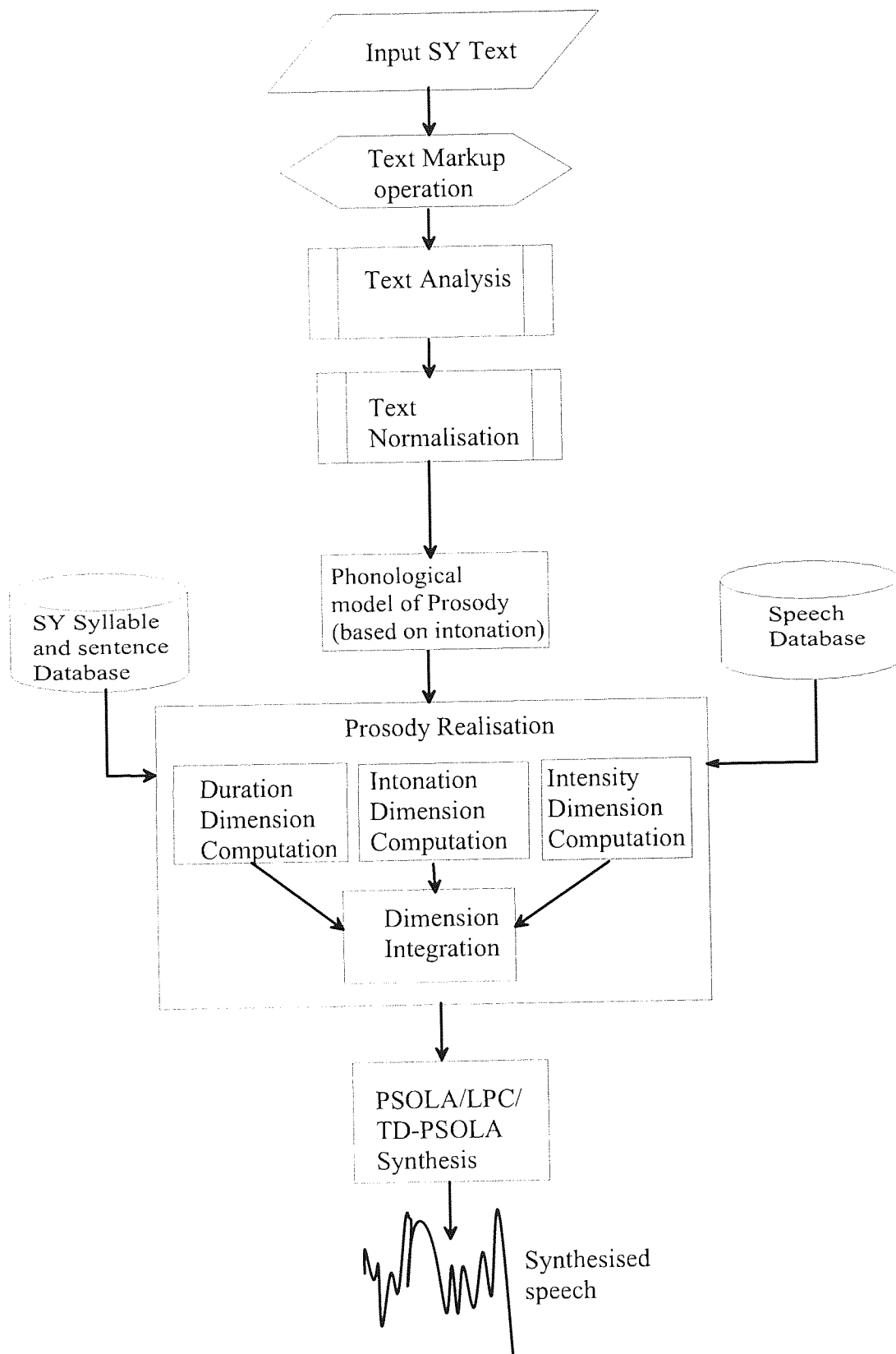


Figure 7.1: SY TTS system overview

and punctuation together with the XML markup.

**Text-to-phonological module:** The text-to-phonological abstraction module generates an abstract representation of the intonation contour, based on the tone phonology of the target language. The abstract structure generated acts as a shell around which other aspects of the speech prosody is developed. We have identified three dimensions of prosody: intonation, duration and intensity. Intuitively, the duration dimension is the most appropriate for organising and structuring the other dimensions of prosody. This is because the timing information is crucial to speech perception and the alignment of intonation pattern as well as ordering of speech waveform (*Lindau, 1986*). We, however, decide to model the abstract structure of speech prosody in our model around the intonation dimension for the following reasons:

1. The basic elements for generating the intonation contour of an utterance, i.e. the tone sequence, can be easily extracted from the text. Also, the intonation events are unambiguously represented in a markup text. These two pieces of information are in closer proximity to the intonation dimension than to the other two dimensions of prosody.
2. The tone phonological rules for generating the abstract intonation contour of an utterance from a sequence of tones have been studied and are well documented for most tone languages. These kind of rules are not available for the duration and intensity dimension of prosody.
3. There are a lot of readily available data on the tonal and intonational dimension of speech prosody. Apart from the fact that the data is easily generated, it is also easy to relate them to acoustic signals in the speech waveform. For example, the  $f_0$  contour of an utterance is widely used to describe intonation phenomena such as downstepping. This type of description is usually not provided in the same detail and clarity for the duration or intensity dimension.
4. It is easy to relate intonation phenomena at the abstract level to the perception of speech sound. This provides powerful evaluation data that can be exploited to stylise and standardise  $f_0$  curves and their subsequent combination to generate  $f_0$  contours.

**Prosody realisation:** The function of the prosody realisation modules is to compute the numerical values for each of the three dimensions of speech prosody. The computation of each dimension is based on a set of rules which establishes the transformation from a canonical to a contextual syllable. The data computed

for each of the dimensions are viewed as the numerical transformation of the abstract model of the prosody. After all the dimensions of interest are computed they can be combined within a single data structure established by the abstract model. The data structure describes the acoustic evolution of each of the prosody dimensions and serves as the input into the speech synthesis module. The acoustic characteristics of canonical syllables, i.e. syllables uttered in isolation, differs from their contextual characteristics due to a number of changes they undergo as a result of the *co-articulation* and *tone-sandhi* phenomena. These changes in characteristics do not inhibit their perceptual identity because syllables can still be recognised, irrespective of the complexity of the context environment.

**Speech synthesis:** The data structure generated by the prosody realisation process is used to synthesise the corresponding speech sound using signal processing techniques such as Linear Predictive Coding (LPC), Pitch Synchronous Overlap (PSOLA) and Time domain-PSOLA (TD-PSOLA). The speech sound generated by the speech signal processor is finally transmitted to the computer speaker.

The major concern of this research is prosody modelling which comprises: (i) text-to-phonological abstraction and (ii) prosody realisation.

## 7.2 Motivation of our approach to modelling

Based on the tone carried by each syllable, it is possible to describe the abstract intonation contour using the tone phonology of the target language. For example, in line with the declination phenomena (*Láníran and Clements, 2003*), when two H-tone syllables follow each other in SY speech, the  $f_0$  curve of the second H tone is realised at a lower peak than that of the first. These rules of tone phonology for generating an abstract intonation pattern are discrete and finite. It is possible, therefore, to formulate an algorithm, based on finite algebra, for representing, manipulating and constructing the intonation pattern of an utterance. Graph, automata, rewrite rules, trees and other tools of discrete mathematics are commonly used in linguistics to implement such a model (*Kornai, 1993*). We have chosen a tree-based approach in this work.

We assume that a number of factors contribute to the perception of smooth concatenation points. These include consistency of  $f_0$  and intensity as well as spectral similarity across the concatenation point. As a first step in the approximation of the  $f_0$  contour, the phonological representation can be conceived as a linear string of tones. Each tone is endowed with constituent structure which can be represented by a number of parameters that can be symbolically related to the perceptually significant points on the  $f_0$  contour. Such symbolic representation, can be modelled as phonological units, whose justification can be found in the regularities of the sound pattern. The stylisation and standardisation techniques are used to achieve this representation in TTS (*d'Alessandro and Mertens, 1995*).

An important feature of the prosodic attributes of speech, however, is that they cannot be defined in absolute terms. For example, there are no absolute phonological statement like “High tone is 300Hz” or “final lengthening increases the duration of the last syllable by 50ms”. This shows that, ambiguity and uncertainty are inherent in the relationship between the linguistic and acoustic levels of speech prosody. We therefore need an effective computational mechanism which is able to account for the emergence of discrete perceptual unit from acoustic continuum. This motivated the use of fuzzy logic based techniques as it has been proven to be effective in the modelling of ambiguous and uncertain concepts (*Zadeh, 1972*). In our model, the relation between continuous phonetic and discrete phonological categories is captured by fuzzy rules. The rule parameters are fine-tuned by subjecting the resulting prosody representation to perceptual evaluation and the parameters adjusted in an iterative manner to improve the perceptual quality.

### 7.3 Overview of the synthesis strategy

Acoustically, speech sound may be regarded as the simultaneous and sequential combinations of pulse, periodic and aperiodic forms of waveform and energy, interrupted by silence and varying duration. Depending on the rate of speech desired, we can generate a range of intervocalic durations by varying the degree of overlap between the two portion of syllables, i.e. onset and rhyme (*Coleman, 1992*). In SY, we assume that the



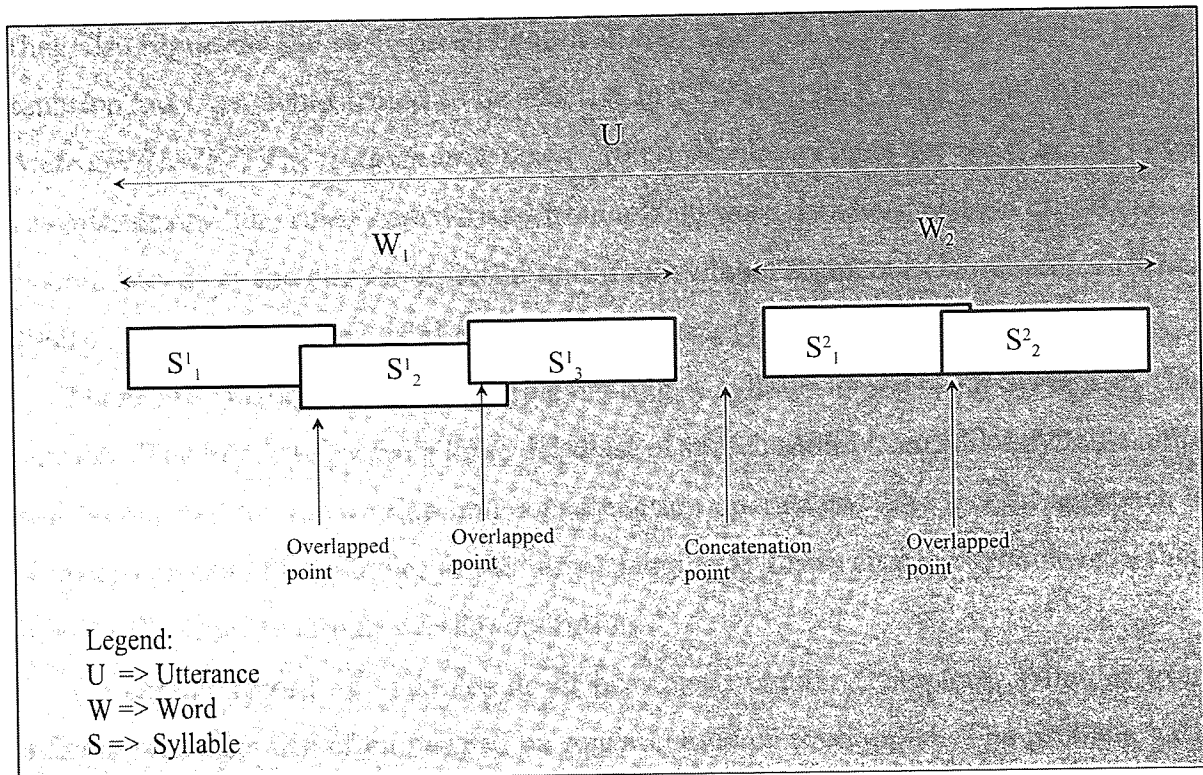


Figure 7.2: Speech synthesis strategy

degree of overlap can be controlled by the duration computed for each syllable with respect to its canonical duration. The degree of overlap between syllables in a word is, in part, determined by the phonological identity of the syllable segments as well as the linguistic context of the syllable in question (*Dilley et al.*, 2005).

Our speech synthesis method employs syllable overlapping (*Coleman*, 1992) and word concatenation strategies (see Figure 7.2). The overlapping strategy is used for joining syllables together to synthesise polysyllabic words. To synthesise utterances longer than one word, each word is first synthesised using the syllable overlapping strategy (*Coleman*, 1992). The synthesised words are then concatenated (*Black and Taylor*, 1997). Pauses of different durations are inserted between words based on a number of contextual linguistic factors such as their proximity to syntactic boundaries, e.g. phrase and sentence boundaries. Afterwards, the overlapped syllables and concatenated words are synchronised and their boundaries smoothed by applying a signal processing technique, such as PSOLA (*Moulines and Charpentier*, 1990).

The syllable based approach to speech synthesis adopted in this work is motivated by the fact that syllables create perceptually and acoustically coherent units and that

they also represent the basic prosodic unit in SY. The idea of word and syllable concatenation has been demonstrated to be effective in TTS speech synthesis. For example, in *MeteoSPRUCE* TTS, *Tatham and Lewis (1999)* developed rules for word and syllable concatenation. The rules were derived from a 2000 word database using a classification system based on the classes of initial and final syllable segments (*Lewis and Tatham, 1999*). A similar approach is also used in ProSynth (*Ogden et al., 1999*)

There are two approaches to implementing a syllable based TTS system in this context. The first is to record a large database of speech sound from which syllable units will be selected for the overlap and concatenation operations. The second approach is to record all possible syllables in the target language. These citation syllables are then used in the concatenate or overlap operation.

We cannot record all syllables in all possible contexts. Therefore, syllables from different contexts will often need to be concatenated or overlapped. The problem with overlapping syllables from different contexts are:

1. the co-articulation of the adjacent segments can lead to significant waveform distortion,
2. the timing of the syllable can be wrong,
3. the linguistic context, such as proximity to sentence and phrase boundaries induces prosodic properties which are difficult to model and modify computationally, and
4. the size of the database limits the possible combinations of syllables that may occur and which may be used in deriving and synthesising appropriate utterances.

We adopt the isolated syllable approach in this work because the acoustic parameters of isolated syllables are stable enough for computational modelling. The stability stems from the small number of contextual factors to be accounted for in their modelling. For example, isolated syllables do not suffer from co-articulation with adjacent syllables. At the same time, the pattern of the acoustic transformation as a result of co-articulation and related prosody phenomena, can easily be studied and brought to bear on an isolated syllable when it occurs in context. This has been demonstrated for Cantonese Chinese (*Lee, 2004*).

The selection of this approach is further informed by our observation of SY utterances which suggests that syllables blend into one another within a word and the

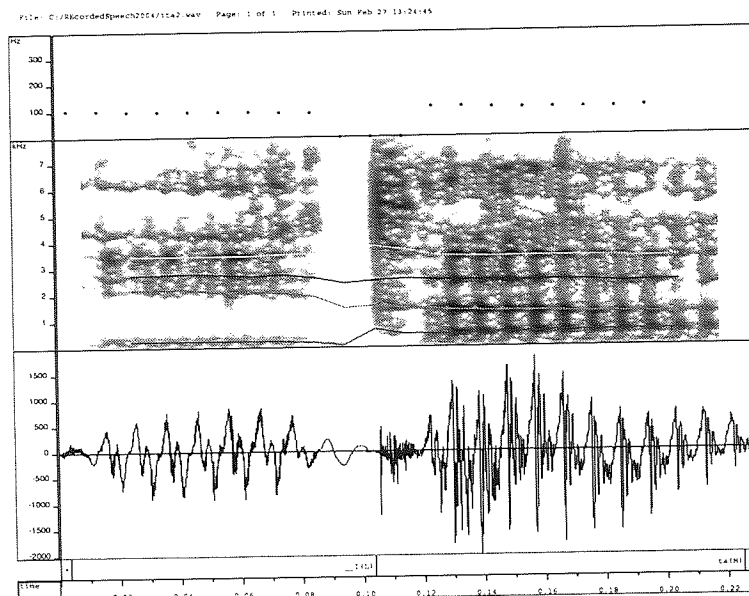


Figure 7.3: Waveform and Spectrogram of SY word “*ita*” (meaning “*Outside*”)

melody of words blend into each other across word boundaries to produce natural sounding speech. It is difficult to describe the acoustic features of speech by using a sequence of sounds. This is because the physical aspects of sound are determined to some extent by neighbouring sounds (*Passonneau and Litman, 1996; Yang, 1998; Gandour et al., 1999*). This leads to variability in the structure and acoustic characteristics of citation syllables when such syllables occur in context. This variability is a systematic variation in spatial structure of the speech waveform. Therefore, it cannot be explained in terms of the properties of the speech production mechanism itself because it has a linguistic origin.

For example, consider the synthesis of the SY word “*ita*” (meaning “*outside*”) which is made up of two syllables “*i*” and “*ta*”. “*i*” is a V-typed syllable carrying a Low tone and “*ta*” is a CV syllable carrying a Mid tone. Figure 7.3 depicts the acoustic characteristics of the word pronounced by an adult male native speaker. The topmost panel of Figure 7.3 is the  $f_0$  pattern on the word. The next two panels are the spectrogram and waveform respectively. The last two narrow panels are the text annotation and duration panels respectively. It will be observed on the speech spectrogram that the articulation of the onset “*t*” of the syllable “*ta*” begins before the vowel “*i*” has been totally released. In this way, the rhythm of the syllable “*i*” flows into that of the onset of syllable “*ta*”.

### 7.3.1 The syllable overlay strategy

Syllable overlay is a statement of temporal relationship between the end of onset syllable and the beginning of the next. The overlapping nature of speech constituents have been noted by many phoneticians (*Lieberman*, 1970; *Local*, 1992; *Coleman*, 1994). A low level of overlay results in a comparatively short proportion of the resulting acoustic segments to being attributed to the coda of the preceding syllable (*Coleman*, 1994). A high level of overlay corresponds to a situation in which less of the coda is masked and the duration of the resulting acoustic segment is longer. Overlay is sensitive to the constituents of both the onset and rhythm.

In speech synthesis, overlapping is one of the principle means by which synthetic speech can be made to sound more connected and less disjointed. At the syllable level, we regard the syllable onset to be overlaid on the nucleus and the nucleus overlaid on the coda, rather than concatenated, as in segmental speech synthesis. This organisation assists in modelling co-articulation. On the duration dimension we assume that compressing a syllable affects the duration of the rhyme primarily, but leaves the onset unaffected. In the synthesis of a sequence of syllables, we assume that syllables overlap to different degrees depending on a number of factors, such as, their phonological attributes and the rate of speech. These assumptions follows (*Coleman*, 1994).

*Wu and Chen* (2001) identified three types of inter-syllabic concatenations inherent in the phonetic properties of phonemes in Mandarin. In *loose concatenation*, the syllables do not have much influence on each other, so co-articulation, if present, is not considered to be perceptually significant. Such syllables are considered to be concatenated with short silence in between. In *overlap* and *tight concatenation*, respectively, strong and medium co-articulatory effects are present. For example, two syllables with overlapped concatenation blended together and hence it is very difficult to obtain a precise syllable boundary.

Our model of temporal compression allows the statement of relationships between syllables in different contexts to be assigned a duration based on the combination of linguistic factors operating in that context as well as the specific attributes of that syllable and its neighbouring syllables. Depending on the degree of overlap between the rhythm and the onset of successive syllables, we can generate a range of durations.

In order to align the  $f_0$  with duration, the degree of overlap is controlled by varying the time between the peak and valley of successive  $f_0$  curves.

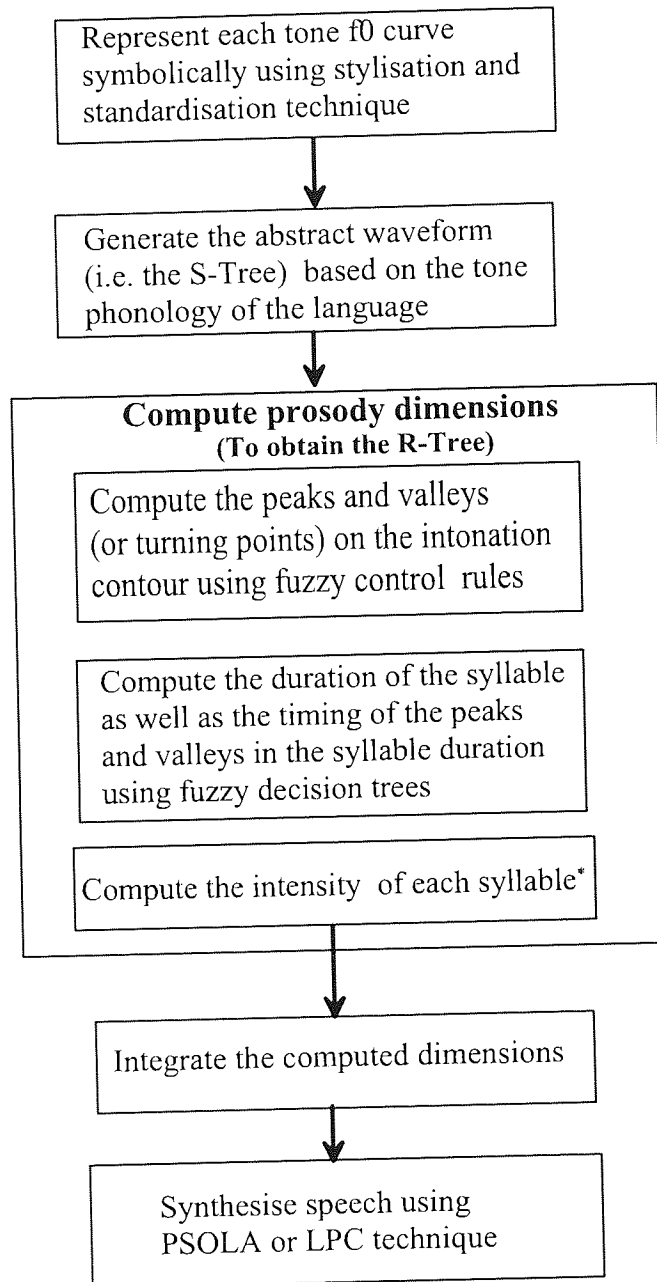
## 7.4 Overview of our prosody modelling technique

There are two goals that must be achieved for our prosody modelling technique to be successful. First, our modelling approach must account for the holistic properties of the prosodic units that are larger than the syllable based on the specific properties of the component syllables. Second, we need to coordinate the temporal events at various linguistic (i.e. syllable and word) levels as well as at the syntactic boundaries within an utterance.

Conceptually, the first goal subsumes the second in that an holistic framework within which linguistic entities can be overlapped or concatenated must incorporate a mechanism by which the overlapping and concatenation operations can be represented. Therefore, the model can be implemented using a modular holistic view of speech prosody.

We propose to implement such a holistic view in a tree based framework using the Relational Tree (R-Tree) technique (*Ehrich and Foith, 1976*). R-Tree is a technique for representing a waveform using a binary tree data structure. In this representation, a succession of peaks and valleys on the waveform are represented as elements in the nodes of the tree. The self-embedding relative heights and depths of peaks and valleys in the tree represent the spatial structure of the waveform.

The construction of an R-Tree involves generating a Skeletal Tree (S-Tree) which represents the abstract structure of the waveform. In our model, the S-Tree for the intonation contour of an utterance is generated using tone phonological rules. The dimensions of the perceptually significant points, corresponding to the peaks and valleys, on the S-Tree are then computed to synthesise the intonation of the target utterance. A complete R-Tree, therefore, contains nodes representing all the phonologically significant peaks and valleys on a waveform as well as their numerical dimensions. The stages in our proposed prosody modelling technique are depicted in Figure 7.4.



\*Not implemented

Figure 7.4: Stages in our proposed prosody modelling

## 7.5 Abstract waveform generation

We applied the following steps to generate the S-Tree in our model:

**Stylisation:** Approximate the  $f_0$  curve of each tone by a simpler function using stylisation technique.

**Standardisation:** Represent each stylised tone type in terms of peak and valley (Collier, 1990).

**S-Tree generation:** Derive an algorithm for computing successive peaks and valleys from a sequence of tones based on the tone phonology of the language.

We have chosen to implement our stylisation from first principles. By this we mean we directly approximate the  $f_0$  data of the voiced portion of a syllable using various interpolation polynomials. These approximations are then subjected to perceptual evaluation in order to determine the one that produces the best speech quality while at the same time facilitating a transparent representation of the  $f_0$  curve.

For the standardisation, the peaks and valleys of the stylised  $f_0$  curve are taken to be the location of the most important phonological events. We then used tone phonological rules to predict the structure and the course of the intonation pattern based on peak/valley data of the sequence of syllables in an utterance. In essence, the tone phonological rules are used to generate the skeletal tree.

As an illustration, let us take the example phonological rules in Figure 7.5(a). These rules specify that if two tones of the same type follows each other in an utterance, the second tone is realised at a lower  $f_0$  peak than the first. Generally, H tones will have higher peaks than the M tones which, in turn, have higher peaks than the L tones.

Given an utterance made up of four syllables with the tone sequence, HLHH, as shown in Figure 7.5(b). We can generate the first level of the S-Tree in Figure 7.5(c) by noting, based on the phonological rules, that the deepest valley in the utterance will be associated with L tone, i.e.  $V_2$ .

Using the phonological rules, the highest peaks to the left and right of the  $V_2$  are  $P_1$  and  $P_3$  respectively. The double circle, used to represent  $P_1$  and  $P_3$  in Figure 7.5(c), indicates that they still dominate other peaks in the waveform representation. The

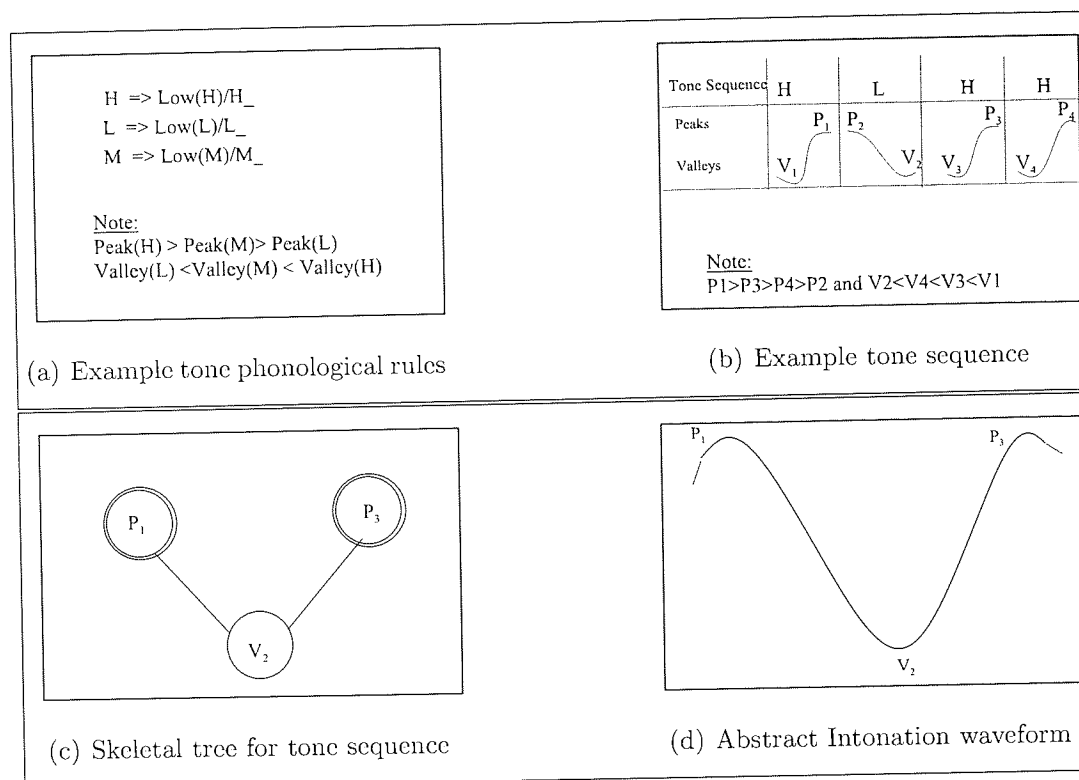


Figure 7.5: Illustration of our prosody modelling technique

abstract waveform for this partial S-Tree representation is shown in Figure 7.5(d). The above discussed process of determining the peak and valley can be applied recursively until all the peaks and valleys on the waveform has been represented by nodes on the S-Tree.

## 7.6 Dimension computation

The S-Tree is converted into a complete R-Tree by computing the respective dimensions, i.e.  $f_0$ , duration and intensity of the peaks and valleys represented in the S-Tree. In this work we only examined the intonation and duration dimensions. An overview of the process for generating the complete R-Tree is shown in Figure 7.6.

### 7.6.1 Computing the intonation dimension

Several requirements have been proposed for the modelling and realisation of accurate intonation. *Crystal* (1969) proposed six requirements for generating an ideal intonation: (i) high accuracy, (ii) high consistency, (iii) automatic applicability, (iv) simplicity of



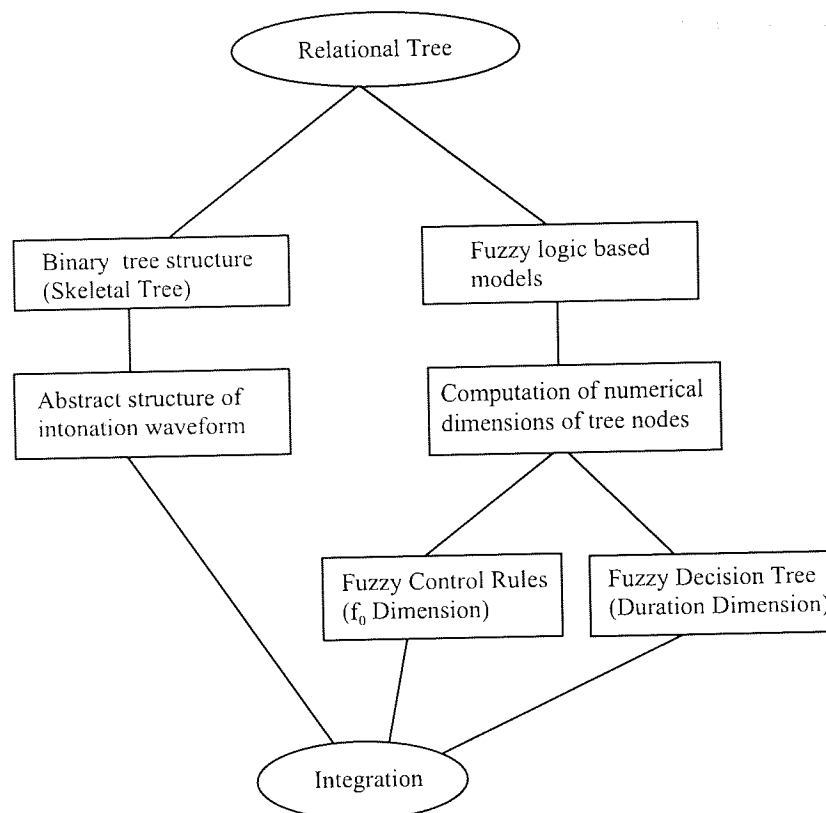


Figure 7.6: Overview of our R-Tree generation technique

symbolic set, (v) a degree of complexity for the symbols which reflects the significant difference in data, and (vi) a restriction to cover only those aspects of intonation which are linguistically significant. *Taylor (2000b)* considered three requirements to be desirable in intonation modelling. These are:

- constrained: the representation should be compact and contain minimal redundancy,
- wide coverage: the model should be able to distinctively describe various intonation patterns,
- linguistically meaningful: the model parameter should be predictable from high-level linguistic information.

*Hirst et al. (2000)* identified four levels for an intonation model to satisfy the global theory of intonation: (i) the physical level which is composed of the acoustic and physiological features (ii) phonetic level, (iii) surface phonology level, and (iv) deep phonology level.

It is clear from the above requirements and criteria for intonation modelling that, to implement an effective intonation model, there is the need to represent intonation

phenomena at various levels. In the context of TTS applications, these levels of representations must facilitate a mapping from more abstract levels (i.e. symbolic) to more concrete levels (i.e. acoustic or numeric). An additional requirement for a TTS application is that it should be possible to relate the parameters used to represent the intonation phenomena to the perception of speech signal.

These requirements and criteria informed our application of the fuzzy control rules in intonation dimension realisation. The linguistic information are the premise of the rule while the consequence is an equation which computes a numerical value. The fuzzy control therefore implements a compositional model of intonation realisation in which complex intonation phenomena, described linguistically, culminate on individual peaks and valleys of syllables to produce the intonation contour. This approach allows us to theoretically bridge the gap between the abstract representation of intonation and the realised numerical value of targets. The implementation of our fuzzy control rule is presented in Chapter 10.

### 7.6.2 Computing the duration dimension

Within the context of the R-Tree based holistic prosody modelling framework, the general goal of our duration modelling is to find a computational relation between a set of factors affecting duration. To do this successfully, the process for computing duration of speech segments must account for all possible factors affecting duration which include the position of the syllable within a word, the context within which the syllable occurs, etc.

An important question is how do we time the peaks and valleys on the intonation contour in order to associate and align them with syllabic position parameter? The association process should describe the structural relationship between intonation stream and the segment stream by specifying which units in one are link to units in the other. The alignment process should describe the temporal relationship between units, and can be important in distinguishing pitch accent type (*Ladd, 1996*).

*Dilley et al.* (2005) have suggested that the timing of the  $f_0$  peak and valley is important in signalling lexical semantic distinctions and that such points are timed consistently with respect to the segmental string. They also argued that there is no

fixed time interval between tones or more generally that the  $f_0$  movements have fixed duration. *Ladd* (2000) has shown that the evidence concerning alignment of  $f_0$  peaks and valleys is relevant to evaluating phonological model of intonation. For example, it has been shown that the two tones in a bi-tonal  $L + H^*$  pitch accent are independently aligned with respect to the segmental string, and not with respect to each other (*Dilley et al.*, 2005).

We have observed similar phenomena in our speech database in that the  $f_0$  patterns of tones align with their associated syllables. The pattern we observed in our data also confirm earlier result from *Xu* (1999a) on Mandarin Chinese. *Xu* shows that as the duration of the syllable changes, the  $f_0$  pattern of a tone tends to move in synchrony with rhyme rather than with the onset. In respect of tone alignment in Mandarin *Xu* (1999a), experimenting with short declarative sentences, states that:

*“ These alignment patterns were interpreted as further evidence that tones in Mandarin are implemented synchronously with the associated syllables and realized in such a way that their target contours are best approximated by the end of the syllable”* (*Xu*, 1999a).

The end of syllable referred to above approximately corresponds to the rhyme in SY syllables.

Based on the above discussion, the duration of each syllable is determined by computing three parameters: (i) time of the peak, (ii) time of the valley, and (iii) the span in time, of the voiced portion of the syllable. In this computation, we assume that the onset suffers little or no time variations. We applied a Fuzzy Decision Tree (FDT) technique (*Janikow*, 1993; *Olaru and Wehenkel*, 2003) to compute the values of these parameters for each syllable in an utterance. Each computed value represents the scale of increase/decrease in the duration of a syllable in a particular context relative to its isolated (or citation) duration. The structure and performance of the FDT is also compared with the Clarification and Regression Tree (CART) method. The implementation of these techniques in the context of our duration model is presented in Chapter 11.

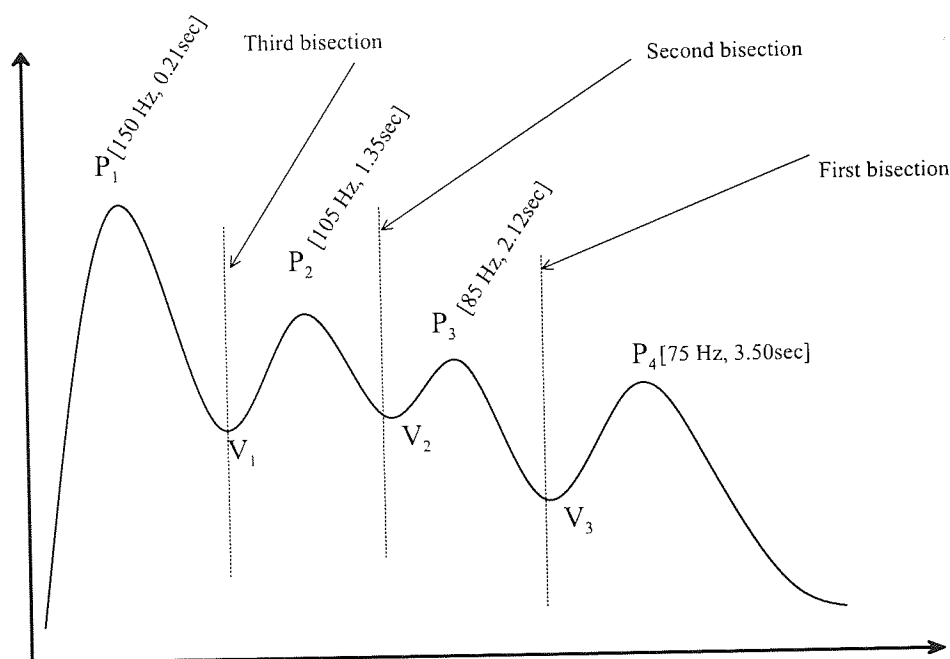


Figure 7.7: Example waveform

## 7.7 Integration of the dimension

Each of the computed dimensions are used to update the S-Tree by adding the computed numerical values to their corresponding nodes. When all the dimensions of interests have been computed and added to the tree, the result is a complete R-Tree. The integration of the abstract and computed values of the prosody dimension provides a comprehensive view of the acoustic attributes of the speech prosody. The values on the  $f_0$  dimension are interpolated to realise the  $f_0$  contour. The peaks and valleys on the interpolated  $f_0$  contour are aligned using the data on the duration dimension.

Using the waveform in Figure 7.7 as an illustration, the R-Tree that will result after the three dimensions of the waveform have been computed is shown in Figure 7.8. It is important to note that the computed values might cause the final waveform to differ (in shape, mostly slightly) from the estimated abstract waveform.

## 7.8 Summary

In this chapter, we have provided an overview of the modelling concepts and design approach to be used in our prosody modelling. The structure of the TTS system,

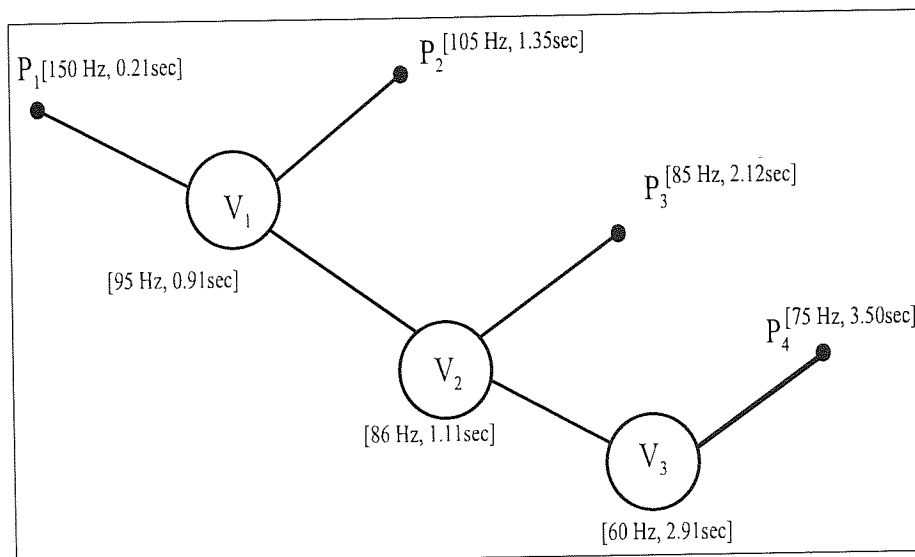


Figure 7.8: R-Tree of the waveform in Figure 7.7

within which the prosody model will be implemented, was presented. The motivation for using the syllable overlapping and word concatenation strategy was discussed. We have illustrated the general framework for implementing our modular holistic model of prosody, based on the Relational Tree technique. The techniques that will be applied to computing the  $f_0$  and duration dimensions, i.e. fuzzy logic and fuzzy decision trees, were also discussed.

# Chapter 8

## Research data

A central issue in the development of a language related technology such as TTS systems is the language resource. This includes specially prepared speech materials, such as text and speech sound, for the target language. The design, collection, recording and annotation of such speech materials depend on several factors. These factors include: the scope of the research, the linguistic features of the language, and the domain of application.

The data used in studies like this varies widely in scale and scope. While some use large, well developed and publicly available databases, such as TIMIT (*Fisher et al.*, 1986), others have developed special database for their research. The specifically developed databases are, in most cases, developed as the research progresses and intended to study or model specific speech phenomena (*Taylor*, 1992; *Vainio*, 2001; *Clark*, 2003).

The creation of an adequate language resource is a very complex task which requires inter-disciplinary efforts amongst phoneticians, linguists as well as the use of powerful speech processing software. There are no publicly available language resources for Standard Yorùbá (SY) so we proceeded to create a small language resource for use in this research.

### 8.1 Collection of text material

The domain for our speech synthesis is in language education and the mass media. We selected four popular SY newspaper and three SY textbooks for creating our text

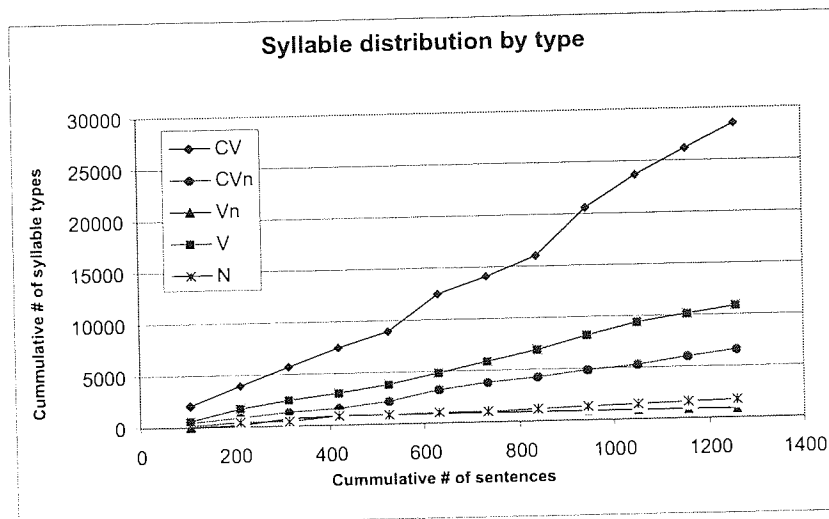


Figure 8.1: Syllable distribution

database. The newspapers are: (i) *Aláròyè*, (ii) *Alálàyè*, (iii) *Ìròyìn Yorùbá*, and (iv) *Akéde Àgbáyé*. All of the newspapers do not use tone marking in their orthography but under-dots are used in *Akéde Àgbáyé* and *Aláròyè*. The texts selected from the newspaper for our text corpus have been tone marked and under-dotted appropriately. Textual anomalies such as numeral, foreign words and proper names are expanded and written in SY orthography using SY accent.

The three textbooks selected are two SY language education textbooks (*Bámgbósé*, 1990; *Owólabí*, 1998) and a book on SY culture (*Ògúnbòwálé*, 1966). In addition to these, we also composed a short SY story and the text is added to the SY text corpus. The purpose of composing the story is to add typical dialogue domain text into the already collected texts. It also allows us to compare the tonal and linguistic distributions in the different domains of SY text.

In the analysis of about 1,500 words in our text corpus, we found that all the SY syllable types are present in the database. Generally, the CV and V type syllables account for more than 75% of the syllable types. While the CVn, N and Vn account for the remaining 25% (see Figure 8.1). Specifically, the CV type syllables are the most commonly occurring and they account for about 43% of the entire database while the V type syllables account for about 32% of the total. The Vn, N and CVn type syllables account for 3%, 10% and 12% of the entire database respectively.

In general, newspaper texts has an average of 2.5 phrases per sentence and bet-

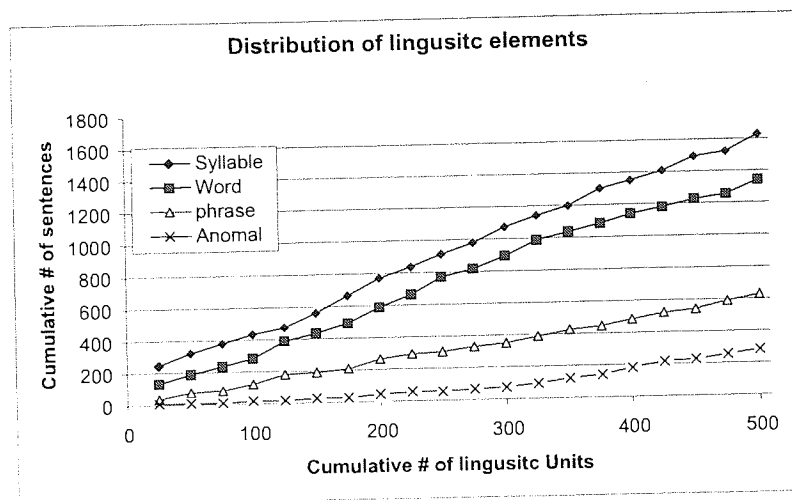


Figure 8.2: Linguistic unit distribution

ween four to fifteen words per phrase. This corresponds approximately to between 15 to 35 syllables per sentence. Textbooks, on the other hand, have an average of 1.5 phrases per sentence and between five and twelve words per phrase. This corresponds to approximately between 10 and 28 syllables per sentence. The number of textual anomalies in the newspaper text (6.4%) is much higher than that in the textbook text (1.4%) (see Figure 8.2). The paragraph lengths in the two domains of text also differ. Generally, newspaper text have more sentences per paragraph than the textbooks.

When compared with other tone languages for which TTS have been developed, e.g. Mandarin, Cantonese and Thai, the tones and syllables of SY are smaller in number and less complex in configuration (see Table 8.1). For example, while SY has just three phonological tones, Mandarin has 4 stable tones and one unstable (neutral) tone (*Wu and Chen, 2001*), Cantonese and Thai have five tones. This implies that the text corpus that covers a relatively larger linguistic elements need not be extremely large or as large as is the case in other languages.

Despite this relatively small text corpus, our text database coverage in terms of syllables and words is rather good for its intended purpose. This result may, however, be subject to changes if a different domain is selected for analysis. For example, the number of words in a phrases and the number of phrase in a sentence depends on the style of writing and the domain of the text.

The text are not syntactically tagged but the linguistic boundaries, such as sentence



Table 8.1: Comparison of syllables and tonal distribution in tone languages

Language	# of Tones	# of Base syllables	Total # of syllables
Mandarin ( <i>Wu and Chen, 2001</i> )	5	408	1313
Cantonese ( <i>Lee et al., 2002a,b</i> )	6	625	1761
Thai ( <i>Thubthong and Kijirikul, 2001</i> )	5	200	1800
Yorùbá	3	230	690

and phrase boundaries, are indicated with appropriate punctuation marks.

## 8.2 Speech corpus

The analysis of the text database discussed above informed the selection of text for our speech corpus. Out of the 690 SY syllables (cf. Chapter 3), we selected 456 isolated syllables. These syllables are carefully selected to reflect the coverage of all syllable types in terms of phonetic and phonological distributions. For example, in the CV syllable type, the manner of articulation of the onset is considered. The onset consonants are selected from each manner of articulation classes, i.e. Stop, Labio-velar, Fricates, Affricate, Sonorant or Semivowel. The selected onset is combined with each vowel type, e.g. Close rounded, Half-closed front, etc., in order to select the syllable for each class of utterance. A list of all the 126 CV type syllables resulting from this selection approach is listed in Appendix A. The same process is repeated for all the syllable types. The data set adequately represents all the five SY syllable types (i.e. *CV*, *CV<sub>n</sub>*, *V<sub>n</sub>*, *V* & *N*).

For our speech database, however, 350 syllables were selected based on the analysis of Yorùbá newspapers and textbooks discussed above. We also generated 360 sentences 95 of which were selected for our prosody modelling. Fifty-five of the sentences are one-phrase sentences while the remaining forty are two phrase sentences (see Table 8.2). All the sentences in our database are semantically well-formed statement sentences and they are selected to reflect common, everyday use of SY. The minimum number of syllables per sentence in our database is 4, with a mean of 6.7 and a maximum of 24

syllables. The H and L tone syllables account for 40% each while the M tone syllables account for the remaining 20%.

Table 8.2: Speech database description

List	Number of syllables	Number of one-phrase sentences	Number of two-phrase sentences
Each speaker	350	55	40
Total	2,100	330	240

### 8.2.1 Recording

The 328 syllables and the 95 sentences were read by three females (identified as f1, f2, and f3) and three males (identified as m1, m2, and m3) all native speakers of SY. The age of the speakers ranges from 21 to 36 years old.

Each speaker read the text at their own pace, resulting in the average number of syllable per seconds ranging from 3.5 to about 6.3. Each recording session lasted for about one hour. The first 15 minutes was to familiarise the speakers with the recording process and environment. All of the speakers participated voluntarily, and were therefore not paid for this activity.

### 8.2.2 Recording equipments and environment

An *ANDREANC* – 61 microphone was used for recording on a Pentium 4.2GHz microcomputer system with on board sound card. The recording took place in a quiet laboratory environment. Two freeware software products were used in this experiment: *Wavesurfer* (Sjolander and Beskow, 2004) and *Praat* (Boersma and Weenink, 2004). *Wavesurfer* was used to record and edit the speech sound as well as to prepare sound files for further processing. *Wavesurfer* was used for this task because it is much smaller and faster. It also provides an easy and intuitive interface for recoding, visualising and editing speech sound. A screen capture of a typical editing session is shown in Figure 8.3.

The parameters for recoding the speech sound is listed in Table 8.3.

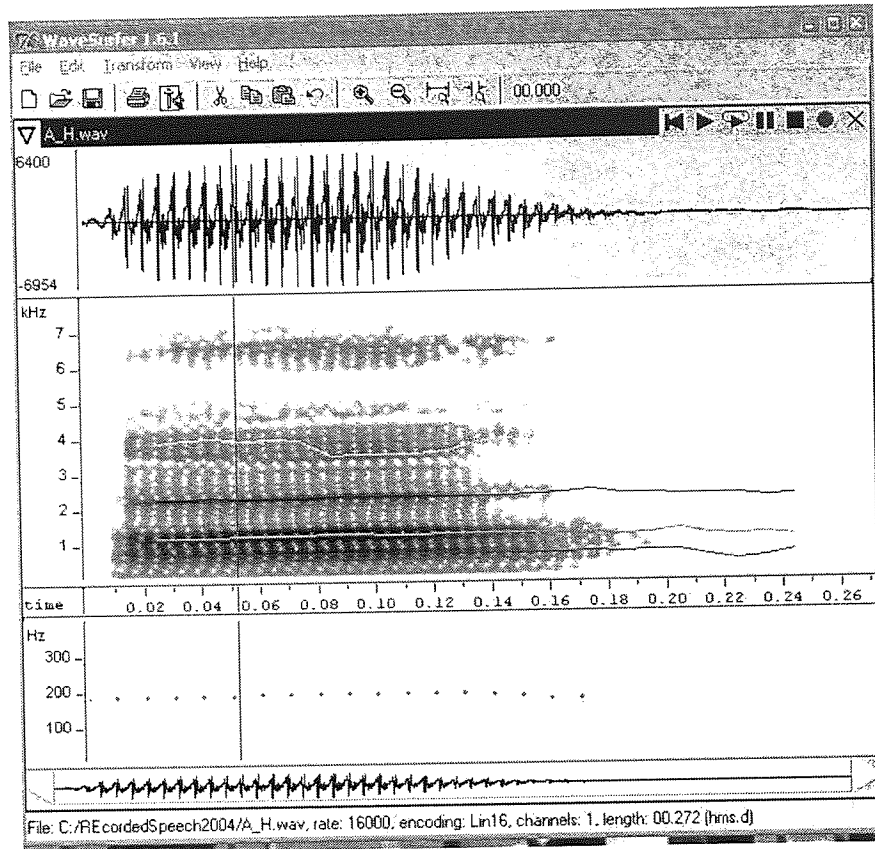


Figure 8.3: Wavesurfer Screen Capture

In order to achieve good recording, recorded speech signals were inspected for the following defects:

- distortion arising from clippings,
- aliasing via spectral analysis,
- noise effects arising from low signal amplitude as a result of quantisation noise or poor signal-to-noise ratio (SNR),
- large amplitude variation, and
- transient contamination (due to extraneous noise).

Recorded speech that has any of the above listed defects is discarded and the sound recording repeated.

### 8.2.3 Speech file annotation

To prepare the recorded waveform for further processing, we replayed each speech sound to ensure that they are correctly pronounced. Artifacts introduced at the end

Table 8.3: Speech sound recording parameter

Serial No.	Parameter	Specification
1	Sampling rate	16KHz
2	Frame size	10ms
3	Window type	Hanning
4	Window length	32ms (512 samples)
5	Window overlap	22ms (352 samples)
6	Analysis	Short time spectrum
7	Set number of channel	1
8	Waveform format	.wav (Microsoft)

Table 8.4: Speech data based annotation symbols

Annotation Level	Annotation symbol
Sentence	Sentence orthography
Sentence boundaries	*
phrase	phrase orthography
phrase boundaries	punctuation marks, e.g. comma(,), full-stop(.) colon(:) etc.
Word	word orthography
Word boundaries	space
Syllable	letter and tone enclosed in bracket, e.g Bá is labelled Ba(H)

and beginning of each recording were edited out. Each speech file is then loaded into *Praat* and annotated manually. To do this, a *Praat* TextGrid is created for each speech waveform file. The TextGrid and waveform files are then loaded for annotation and editing.

There are four tiers in our annotation: sentence, phrase, word, and syllable. Sentences comprising only one phrase do not have phrase tier annotation. The labelling is done to identify syllabic and prosodic constituent boundaries. Fundamental frequency contours (i.e.  $f_0$ ) and formant frequencies (i.e. F1, F2, F3 and F4), were displayed together with the speech waveform. Phrases are separated by a comma in the text and by a pause in the speech. The annotation symbols used are listed in Table 8.4.

For the citation syllable database, the speech waveform boundaries are determined from certain physical characteristics of the speech signal such as regions of discontinuity in excitation, abrupt change in fundamental frequency contour, and visible formant

structure. Regions with very low or no waveform activities correspond to pauses in the utterance.

In the annotation of the syllable speech files, only one tier is specified, i.e. the syllable tier. The symbol \* is used to annotate the syllable boundaries. Each syllable is labelled with its letters, with its associated tone enclosed in parenthesis. The screen capture in Figure 8.4 shows the annotation of the SY syllable “dé”.

When annotating the sentence speech files, the labelling order is: (i) sentence, (ii) phrase (if more than one), (iii) word and (iv) syllable. This labelling order simplifies the detection of boundaries in smaller linguistic units since their annotation is guided by those of the larger ones. For example, after the annotation of the word tier in a two-syllable word, the beginning of the first syllable and the end of the last syllable can be easily determined. This approach also reduces annotation error since larger units are much easier to identify from speech sound signals and perceptually from listening to a replay of the sound segment. The screen capture in Figure 8.5 shows the annotation of the two-phrase SY sentence “Ópé kí ó tó dé, kò tètè lọ” (meaning “[*He late before coming, he not quickly go*] *He came late, and did not go early*”).

Both the spectrogram and the waveform are used to determine syllable and word boundaries. For certain types of syllables in the continuous speech for sentences, we found that boundaries between syllables were hard to determine. This occurred between V-V pairs or between V-CV pairs where C is a semi-vowel such as /y/ or /w/. In this situation, we employed listening tests in addition to speech spectrograph and waveform characteristics for syllable boundary detection. Where the boundaries were in doubt, we found the earliest reasonable position and the latest reasonable position, then placed the boundary half-way in between. For voiced plosives, i.e. /p/, /t/ and /k/, we placed the syllable boundary in the centre of the closure. We note that the voiced plosives show strong segmental effects on the  $f_0$  curves of the syllable in which they occur.

The database comprises a subset of possible linguistic structures, each with a number of exemplars. The database provides us with materials for the analysis of the spectral, temporal and intonational phenomena that we aim to synthesise (*Ogden et al.*, 2000).

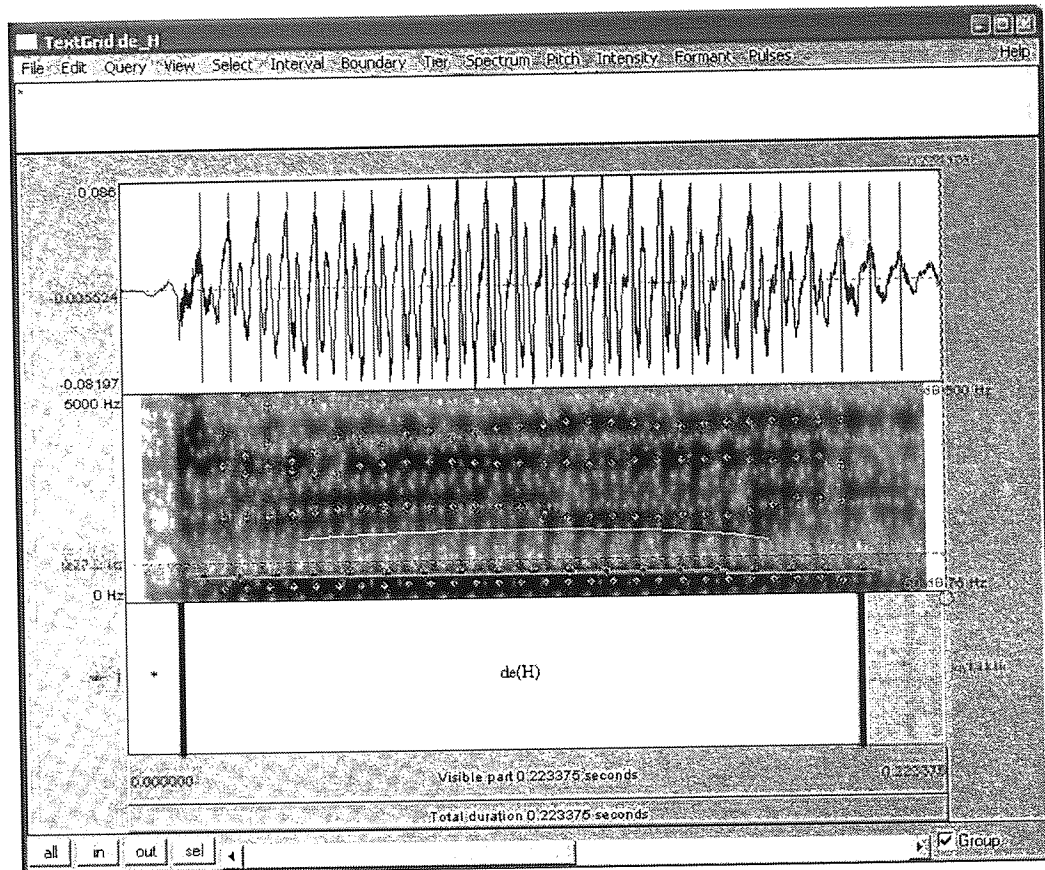


Figure 8.4: Screen capture of annotation of the syllable *dé*

The data for our prosody model were extracted from the annotated speech database. The duration data for the sentence in Figure 8.5 is shown in Table 8.5. The fundamental,  $f_0$ , frequency data are extracted based on the duration data for the linguistic item of interest. A typical  $f_0$  data extraction panel is shown in Figure 8.6.

### 8.3 Summary

In this chapter, we have presented the database that forms the basis of our prosody modelling design and implementation. The data were collected from carefully selected SY sentences from newspapers and textbooks. The speech database contains 95 sentences and about 380 syllables. Each one recorded by six native speakers, three males and three females. The data were annotated using a specially designed annotation scheme (cf. Section 8.2.3).

The amount of data used in our model is relatively small when compared to that

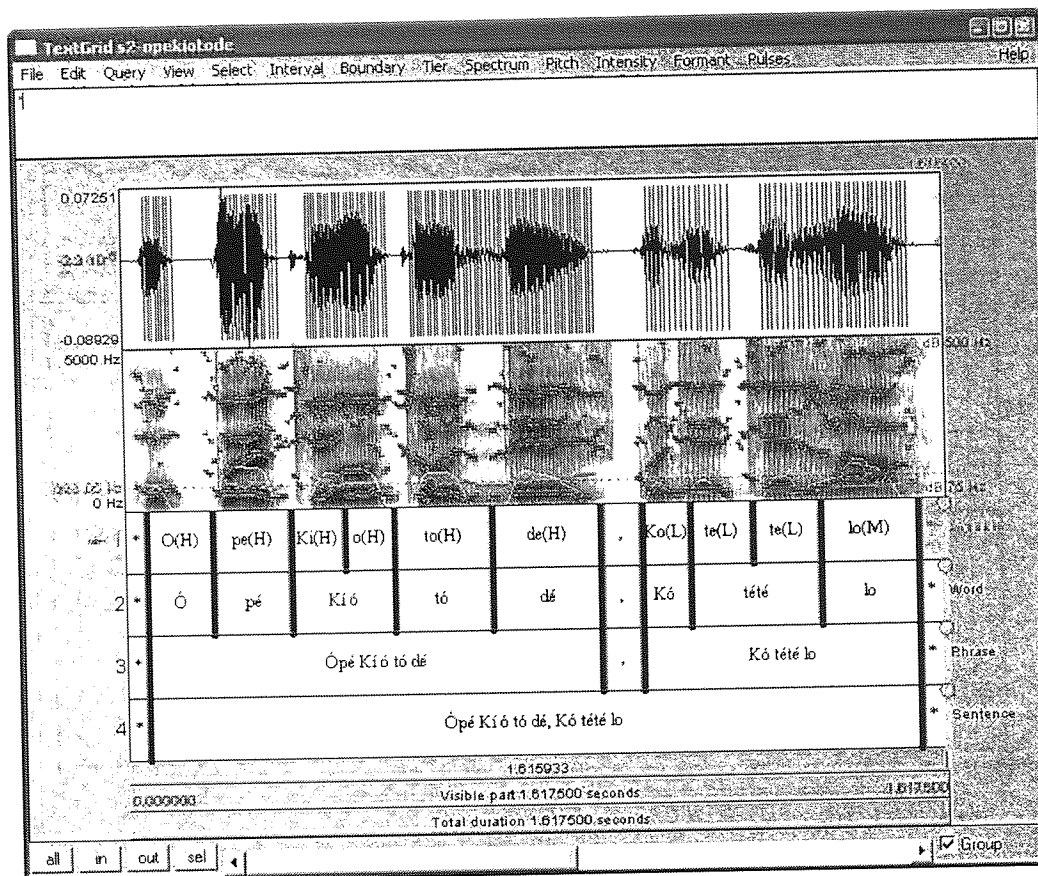


Figure 8.5: Screen capture of annotation of the sentence Ópé kí ó tó dé, kò tètè lo

used in data-driven approaches such as the one reported in *Vainio* (2001). However, our data is much larger than that used in an analytical approach, such as the one presented by *Sun* (2002) who used a database consisting of about 40 minutes of speech read aloud by a female professional announcer, labelled using the ToBI (*Venditti*, 1995; *Campbell*, 1997) system. In total, our database consists of 14,377 syllables. One reason for using this relatively small database is that our approach combines data and expert knowledge in the form of rules. The rules act to augment that data.

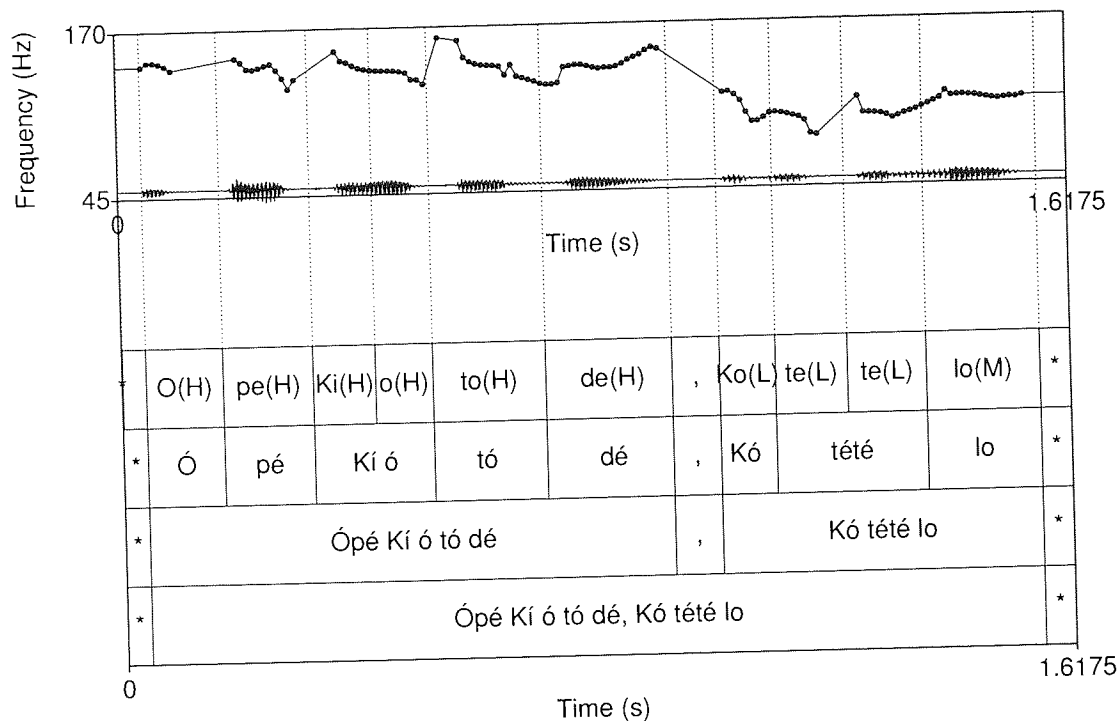


Figure 8.6: Example  $f_0$  data for the sentence “Ópé kí ó tó dé, kò tètè lo”

Table 8.5: Duration data for a sentence

Start time (sec)	End time (sec)	Syllable
0.000000	0.0225000	*
0.0225000	0.1350000	O(H)
0.1350000	0.3372316	pe(H)
0.3372316	0.4260830	ki(H)
0.4260830	0.5553215	o(H)
0.5553215	0.7047534	to(H)
0.7047534	0.9475000	de(H)
0.9475000	0.9834238	,
0.9834238	1.1075000	ko(L)
1.1075000	1.2400000	te(L)
1.2400000	1.3550000	te(L)
1.3550000	1.5775000	lo(M)
1.5775000	1.6251100	*



## Part IV

# Model Implementation

## Chapter 9

# Standard Yorùbá prosody model design

This chapter provides the background to our prosody model design. It contains a description of the domain of prosody in SY as well as a review of stylisation and standardisation techniques used in tone and intonation modelling. This chapter also contains our experiment into the stylisation of SY tones from *first principle* and our standardisation approach. Also documented in this chapter is the design of the algorithm for generating the S-Tree in our R-Tree based prosody modelling.

### 9.1 The domain of prosody in SY speech

The traditional approach for describing prosodic realisation at sentence level usually implies linear construction such as declination lines (*Pierrehumbert, 1981; Thorsen, 1986*) or tonal grids (*Gårding, 1983*). These rigid templates have been found unsuitable for modelling intonation and more flexible alternative approaches have been proposed. An alternative model, the hierarchical organisation model, interprets an  $f_0$  contour as a complex pattern resulting from the superposition of several components. In this model,  $f_0$  curve is often considered as the superposition of relatively fast pitch movements on a slowly declining line.

An example of this kind of model (*Kohler, 1997a,b*) describes the downstepping and reset model in which the amount of downstepping in an utterance and the reset at

syntactic boundaries can be controlled. Also the Fujisaki (*Fujisaki and Hirose, 1982*) model allows for the generation of sentence  $f_0$  contour using response filter models. The hierarchical model allows some analytical separation of the model components. This helps to decide under which conditions and to what extent the concrete shape of the  $f_0$  contour is determined by linguistic features, such as the lexical tone, and non-linguistic factors, such as intrinsic and co-articulatory  $f_0$  variations and speaker characteristics.

None of the above sentence contour models alone are sufficient to precisely characterise the intonation prototypes of SY discussed in *Connell and Ladd (1990)* and *Láníran and Clements (2003)*. SY intonation cannot be modelled with the single declination, or rising, lines as suggested by the sequential model (*Pierrehumbert, 1981*) because SY intonation is characterised by a host of local events, such as the *tone sandhi* phenomena, which cannot be accounted for using a single predetermined line. Whereas declination is a global feature of the intonation contour that seems to be programmed over a domain encompassing a major syntactic constituent, the fall-rise patterns in SY intonation appear locally at the site of a variety of syntactic boundaries and are superimposed on the more global declination function. Moreover, observations in our data suggest that the degree of fall or rise exhibited by the fall-rise pattern in different sentences varies.

It is not possible to adequately model SY sentence intonation with a single sentence component of response filter alone as suggested by the Fujisaki model, because sentence command generates slow melodic movement making intonation events such as *H-rising* and *L-lowering* difficult to model. A way to incorporate this into the Fujisaki model is to add additional phrase commands. However, adding phrase commands will result in the creation of intonation segments which may not correspond to any phrasal boundary.

Based on the weaknesses of the models discussed above, we decided not to use a particular intonation model, but to stylise and standardise the  $f_0$  curves associated with each SY tone with the aim of concatenating them according to SY phonological rules. We assume that prosody can be modelled as the superposition of independent multi-parameter waveforms which belong to hierarchical linguistic levels (*Morlec et al., 2001*). The prototypical structure and movements of each syllable “prosody” parameter

## CHAPTER 9. STANDARD YORÙBÁ PROSODY MODEL DESIGN

is stored in a speech inventory and dynamically used to generate the structure of longer linguistic units under the control of phonological rules. In essence, each syllable participates in the encoding of the abstract linguistic representation as well as the phonetic realisation of the various dimensions of speech prosody.

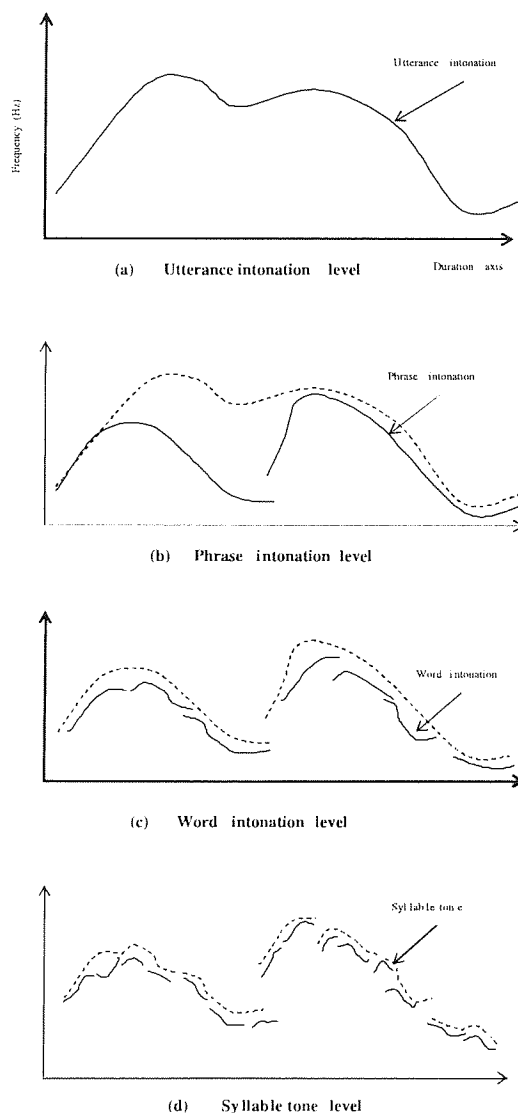


Figure 9.1: Levels in Intonation modelling

The motivation for our approach is that the information about speech prosody, e.g.  $f_0$  variation, that is likely to be relevant to the perception of tones are generally preserved in the recorded speech signal. This information can be captured symbolically and used to represent the pattern of each tone. Linguistically based heuristics can then be applied to generate the abstract phonological structure of an utterance from the tonal elements. The acoustic signal is generated from the phonological structure

via computational mechanisms that are modelled to mimic human expert knowledge in the context of Artificial Intelligence (AI).

The intonation of an SY utterance can be represented as a hierarchical structure with elements exhibiting specific behaviour at different levels in the hierarchy (*O'Shaughnessy and Allen, 1983*)(see Figure 9.1). At the highest level is the sentence intonation pattern. The sentence intonation pattern may contain one or more phrase intonations depending on the number of phrases in the sentence. The phrase intonation pattern of a two phrase structure is shown in Figure 9.1b. Below the phrasal intonation is the word intonation level which represents the  $f_0$  contour of each word (see Figure 9.1c). Below the word intonation level is the syllable tone level. The syllable tone level is depicted by the  $f_0$  curve of each syllable that makes up words in the utterance (see Figure 9.1d). Since the syllable is our unit of speech synthesis, the tone on each of such syllable forms the primitive element that must be represented at the lowest level on the abstract intonation hierarchy.

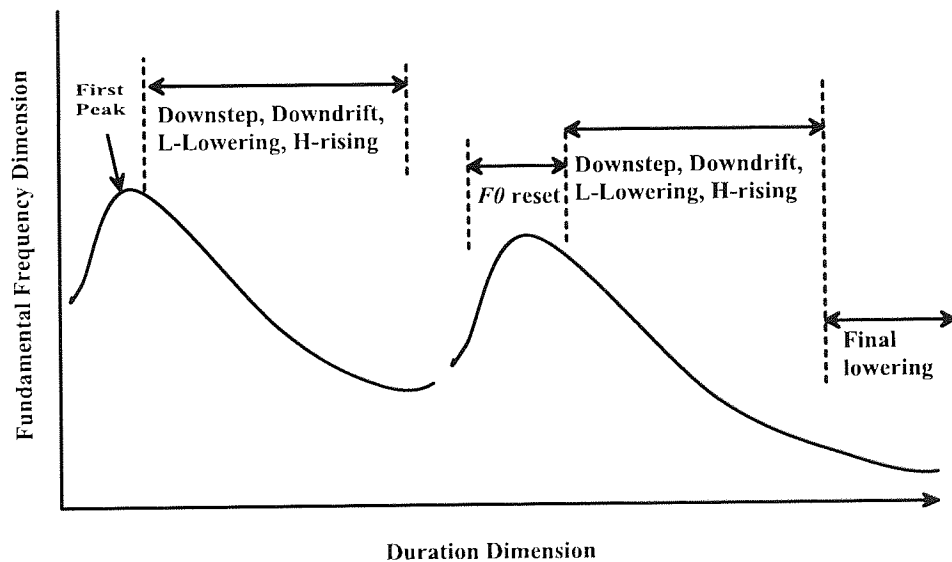


Figure 9.2: Domain of intonation phenomena in Yorùbá utterance

The symbolic representation of these primitives can serve as input to an algorithm which predicts the spatial structure of the intonation corresponding to an utterance based on the underlying intonation phenomena. The domains of some of the SY intonation phenomena are depicted in Figure 9.2 and the way in which their interaction controls the intonation contour generated for an SY utterance has been discussed by

*Connell and Ladd* (1990) and *Lánúran and Clements* (2003).

This chapter is devoted to the discussion of how the basic elements, i.e. tone  $f_0$  curves, are mathematically approximated and symbolically encoded. The design of the algorithm used to generate the abstract intonation pattern from the symbolic representation is also discussed.

## 9.2 $f_0$ stylisation and standardisation

Stylisation is the approximation of acoustic data, extracted from a pre-recorded natural speech signal, by a mathematical function or computational construct (*Hart*, 1991; *d'Alessandro and Mertens*, 1995). The aim in stylisation is to approximate the  $f_0$  curves in the acoustic data with a simpler mathematical function that preserves the perceptual quality associated with the represented data. In intonation modelling, the function generated by the stylisation process forms the basis of a computational model that must be used for generating the  $f_0$  contour that is functionally equivalent to a natural intonation pattern. If the stylised  $f_0$  curve is equivalent to the perceptual impression of its pitch, the standardised symbol based on it can be used to represent, manipulate, generate and communicate the abstract intonation pattern.

In text-to-speech synthesis research, fundamental frequency ( $f_0$ ) has been the subject of many stylisation experiments since it is an important acoustic dimension of prosody around which the other acoustic correlation of prosody can be organised. It is also easier to relate the  $f_0$  dimension to the perceptual quality of speech such as naturalness and intelligibility. Standardisation is the symbolic encoding of stylised  $f_0$  curves, by means of suitable labels, in order to reduce the stylised curve to a sequence of discrete categories corresponding to linguistic entities.

The aim in standardisation is to capture aspects of speech prosody that are assumed to be relevant for language communication using finite set of symbols. The finite discrete symbols represent perceptually motivated tonal movements for each tone types. By reducing the stylised curve into a sequence of discrete categories in this manner, it is possible to generate the symbolic representation of tones and derive the underlying abstract structure of intonation. That structure can be related to a text through sym-

bolic assignment of tones in the text to the identified categories. This is very important for TTS system implementation because it will facilitate the unambiguous extraction of tonal information from text and their subsequent assignment and computational processing.

The following are the requirements for our  $f_0$  stylisation and standardisation process:

1. The representation should be objective, robust and easy to interpret phonologically.
2. It should model the evolution of  $f_0$  over a syllable and it should be capable of being used as the basis for the prediction of intonation phenomena such as declination,  $f_0$  reset, L lowering, H raising, etc.
3. The representation should be quantifiable to enable the subsequent estimation of local and global prosody parameters.
4. The transcription should be theoretically neutral. This will facilitate its application to prosody models irrespective of the theoretical support.
5. The representation should facilitate the time-alignment of  $f_0$  contour to syllable.

The above listed requirements suggest the need to reach a compromise between mathematical accuracy and simplicity on one hand and perceptual quality and ease of implementation on the other hand.

### 9.3 Related works

Stylisation and standardisation have been the focus of a number of studies (*t'Hart and Cohen, 1972; Taylor, 1994; d'Alessandro and Mertens, 1995; Pirker et al., 1997*) and the technique has become established in tone and intonation modelling. The different versions of this approach differ in the mathematical function for approximating the  $f_0$  data of the basic intonation elements and the symbol for representing them. The major focus of most work, however, is the symbolic coding of the  $f_0$  curve in order to produce an inventory that can be related to the linguistic elements and structure of intonation.

Traditionally, mathematical approximations of  $f_0$  contours for stylisation are described in terms of parameters using averaging, additive, multiplicative and/or gradient functions (*Pirker et al., 1997; Campione et al., 2000*). The linear, quadratic and cubic



Figure 9.3: A demonstration of IPO stylisation process (*'t Hart et al. (1990)*)

functions have been proposed and used in stylisation. In this section, we discuss the work reported in the literature in respect of these methods and their implementation.

### 9.3.1 Linear function based model

The IPO's (*Instituut voor Perceptie Onderzoek*) model is perhaps the most popular stylisation technique that employs linear function in the approximation of  $f_0$  data. The IPO's approach is based on the *close-copy* concept. A stylised  $f_0$  contour is a close-copy of the original if it is perceptually identical to the original. To achieve this, native speakers of the target language are asked to judge a recording of an utterance with a synthesised version of the utterance in which the  $f_0$  contour has been replaced by a sequence of straight lines (on a logarithmic scale). Figure 9.3 depicts a typical stylised  $f_0$  contour.

In order to standardise the stylised  $f_0$  curve, the straight lines obtained using this stylisation are replaced by a set of prototype straight lines called the standardised pitch movement. The standardisation process in IPO involves the adaptation of the



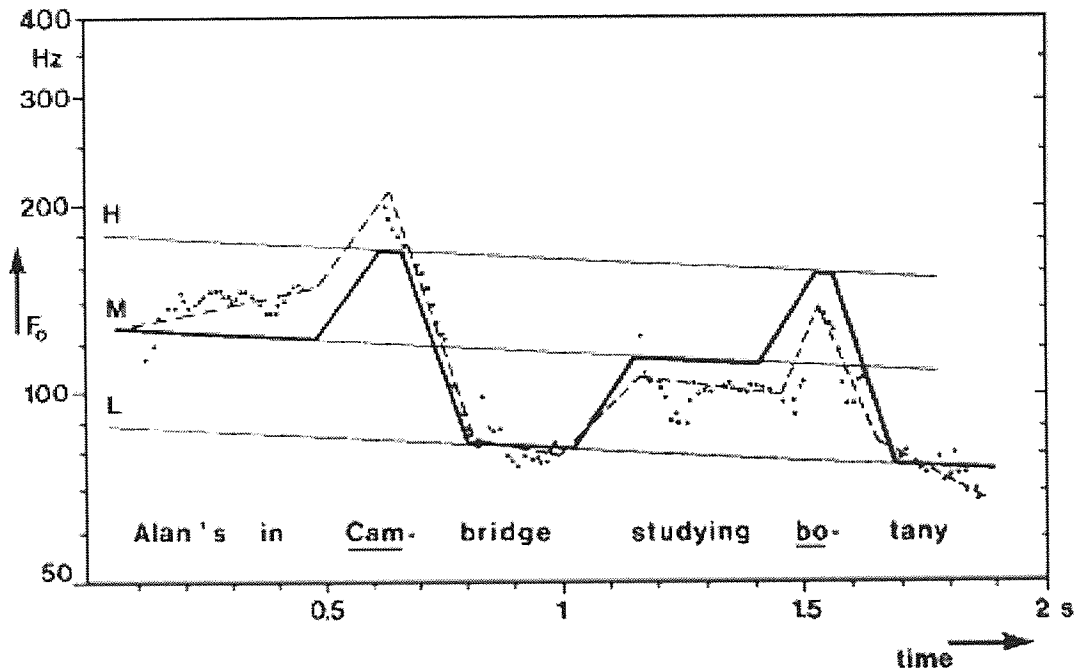


Figure 9.4: IPO standardisation from

stylised contour to a grid of three continuously decreasing lines, i.e. Low (L), Mid (M), High(H). These lines represent the declination lines and are determined based on the criterion of perceptual equivalence to the close-copy representation. In this way, it is possible to describe intonation contours with a series of straight lines falling between the three declination lines (see Figure 9.4). Pitch accents or tones are represented by rises and falls between these declination lines. This procedure facilitates the reduction of information at each stage from the continuously varying  $f_0$  contour up to the standardised intonation patterns.

This approach has been used in the stylisation of the  $f_0$  contours in Dutch ('tHart, 1991; Cohen, 1995). 'tHart (1991) claimed that the linear model produces a close copy that is perceptually indistinguishable from a parabola base model. An advantage of this approach is that it can be adapted to represent the three phonological levels in SY tonology. It is important to note, however, that the intonation contours generated are not generally phonologically motivated since their generation does not depend on phonological rules. Due to the rigid lines used by this model, it will be very difficult to

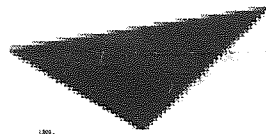
model SY intonation phenomena such as downstep, which is triggered by local context of tone (*Láníran and Clements, 2003*), and not a global declination as suggested by the IPO model.

Other linear function based models have also been proposed and applied to intonation modelling. In the Lund model (*Gårding, 1983*) for example, intonation stylisation is performed by linear interpolation between target values at vowel onset and the end of the syllable. In *Hirst (1992)*, an attempt is made to derive the phonological description of an utterance from its  $f_0$  contour by fitting linear splines to a number of target points.

In *d'Alessandro and Mertens (1995)*, an automatic stylisation model based on linear interpolation was proposed. Syllable pitch-contour are categorised as dynamic or static depending on the presence or absence of a perceptible pitch movement. A straight line is fitted between the pitch values at the start and end of the analysed interval. A turning point can be found at the point of maximum difference between the pitch data and the fitted line. A merge step, which uses the differential glissando threshold (DGT) is then applied (*d'Alessandro and Mertens, 1995*). The  $f_0$  contour is transformed into a sequence of tonal segments which are either static or dynamic based on the glissando threshold which varies with the duration of syllables.

Another semi-automatic stylisation approach, called Prosogram, was proposed by *Mertens (2004)*. The Prosogram stylisation process consists of two parts: (i) the pitch contour and (ii) one or more time-aligned annotations. The phonetic transcription (e.g. IPA, SAMPA) or other types of annotation (e.g. text, tones, stress type) is required for the computation of the stylisation. In Prosogram, stylisation is applied to the  $f_0$  variation of the vocalic nuclei of syllables. The processing steps in Prosogram are (*Mertens, 2004*):

1. Calculate acoustic parameters:  $f_0$ , intensity, voicing ( $V/UV$ ).
2. Obtain a segmentation into vocalic nuclei: use a phonetic alignment, select the vowels, then select the voiced portion of vowel that has sufficient intensity (using difference thresholds relative to peak intensity).
3. Stylise  $f_0$  of vocalic nuclei.
4. Determine pitch range used in speech fragment. Plot stylised pitch and phonetic annotation. Use a musical (semitone) scale and add calibration lines at every 2 semitones for easy interpretation of pitch intervals.



Aston University

Illustration removed for copyright restrictions

Figure 9.5: Early, medial and late peak alignment in KIM ( $V_{on}$  indicates the onset of the stressed vowel) (*Braunschweiler* (2003))

Some other models, such as those proposed by *Pirker et al.* (1997) and *Kohler* (1997a), deal with the translation of phonological representation of intonation into a concrete phonetic model. In such a model, no specific functional approximation of  $f_0$  data is undertaken. Rather parameters are used to specify part of the  $f_0$  contour. In KIM (Kiel Intonation Model) (*Kohler*, 1997a), for example, the fine detail of pitch movements within peaks and valleys are expressed by the division of peak alignment into early, mid and late (see Figure 9.5). The model is composed of two classes of rules: (i) rules defining the phonologically-controlled prosodic pattern by a small number of significant  $f_0$  points, (ii) rules that output continuous  $f_0$  contours based on articulation-related modification. The model also takes *microprosodic phenomena* into account in deciding the output  $f_0$  contour.

In a similar vein, *Taylor* (1994) introduced an intonation modelling technique in which an  $f_0$  curve is analysed as a linear sequence of three primitive elements: rise, fall and connection. Peak accent is modelled by describing the rise and fall part of the accent. Equation (9.1), called the *quadratic monomial* equation, is used to model both rise and fall.

$$y = \begin{cases} 1 - 2x^2, & 0 < x < 0.5 \\ 2(1 - x)^2, & 0.5 < x < 1.0 \end{cases} \quad (9.1)$$

To model the variation in rise and fall, two scaling factors were used to allow the  $f_0$  curve to be stretched or compressed along two dimensions, amplitude and duration, labelled  $x$  and  $y$  respectively. The shapes of the rise and fall elements are given by



Figure 9.6: The quadratic monomial model (source: *Taylor* (1994))

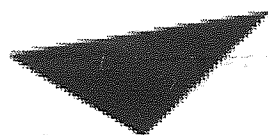
Equation (9.2):

$$f_0 = \begin{cases} A - 2A(t/D)^2, & 0 < t < 2/D \\ A - 2A(1 - t/D)^2, & 2/D < t < D \end{cases} \quad (9.2)$$

where  $A$  and  $D$  are the scaling variables.  $A$  is the element amplitude and  $D$  is the element duration (see Figure 9.6). The connection elements are modelled by using linear interpolation. Although the model gives the rise and fall part of the  $f_0$  in one equation, these parts still have to be modelled separately as there was no observable simple relationship between the amplitude, duration and gradient of a tone.

### 9.3.2 Quadratic model

In order to represent the intonation contour more accurately, a number of approaches involve the application of quadratic interpolation techniques. The most popular amongst these is the MOMEL (MOdélisation de MELodies) (*Hirst*, 1992; *Campione et al.*, 2000) stylisation method. MOMEL uses a quadratic spline function which comprises a sequence of parabolic segments for stylising  $f_0$  curve. Unvoiced segments are interpolated so that the resulting curve presents no discontinuities. This method results in the generation of a continuous, smooth curve which produces a better result than linear



**Aston University**

Illustration removed for copyright restrictions

Figure 9.7: Calculation of a local target point in MOMEL (source: *Campbell (2000)*)

interpolation.

A quadratic equation of the form  $y = a + bx + cx^2$  is used to approximate the  $f_0$  curve within an interval  $\langle t, h \rangle$ . In this equation,  $t = -b/2c$  and  $h = a + bt + ct^2$  (see Figure 9.7). The computed target corresponds to the minimum and maximum of the parabola within the analysis window.

Campione *et al.* claimed that the characteristic of the quadratic spline function are also shared by a more complex stylisation function proposed by *Fujisaki and Hirose (1982)*. However, a quantitative and qualitative evaluation of the MOMEL  $f_0$  stylisation algorithm for French and Italian reveals three types of systematic errors (*Oliver, 2005*): (i) target points in wrong positions, (ii) redundant target points and, (iii) missing target points. As a result of these errors, a final contour can be misrepresented in such a way that a final fall is replaced by a final rise, and, in the second case, utterance initial stylised  $f_0$  values may be too high for a particular speaker.

### 9.3.3 Cubic model

Cubic or third degree polynomials are popularly used in the approximation of  $f_0$  curves for tone languages. *Lindau (1986)* modelled the  $f_0$  curve of Hausa, a two tone language, by interpolating between turning points, using a piecewise application of third-degree polynomials. Each turning point is specified as  $t_n, y_n$ , where  $t_n$  is the turning point time and  $y_n$  is the  $f_0$  at that time. The function  $y(t)$  computes the  $f_0$  at time  $t$ . Each

pair of points are joined by a curve, subject to the condition that the turning points have zero velocity. A program calculates the constants  $a_0, a_1, a_2$  and  $a_3$  in the equation  $y = a_0 + a_1t + a_2t^2 + a_3t^3$ .

Lindau concludes that the fit between the real and the generated curve is *quite good*. Furthermore, Lindau found that the timing of the turning points in the  $f_0$  curve in relation to the segmental structure of an utterance is a necessary part of intonation specification. This is because the maximum and minimum of the  $f_0$  curve do not generally coincide within the same segment. Instead, there is a strong tendency for the turning points in a sentence to occur at, and around, the syllable boundaries.

Similarly, in *Gandour et al.* (1999) and *Thubthong and Kijisirikul* (2001), 3<sup>rd</sup> degree polynomials were shown to be adequate for modelling the  $f_0$  curve for the five tone in Thai language. In *Wang* (2001) and *Wu and Chen* (2001), the first four coefficients ( $a_0, a_1, a_2, a_3$ ) of the discrete Legendre transformation were used to model the  $f_0$  curve of Mandarin Chinese tones. This transformation is basically a 3<sup>rd</sup> degree polynomial. More complicated mathematical formulae such as Bézier curves (*Aguero and Bonafonte*, 2004) have also been proposed and used in  $f_0$  curve approximation. Such models tend to represent both the redundant and perceptible  $f_0$  movements resulting in complex model with large parameters.

### 9.3.4 Comparison of models

The models discussed above show the diversity of the approaches used in stylisation and standardisation. In stylisation, the main bone of contention is the degree of the polynomial for approximating the  $f_0$  data (i.e.  $f_0$  curve or contour). An advantage of linear interpolation approach is that it is fairly simple to implement and requires fewer parameters, resulting in a simpler and faster  $f_0$  generation. *tHart* (1991) argued that stylisation by curvilinear functions is not perceptually distinguishable from that using straight-lines.

But *Campione et al.* (1997) demonstrated that stylisation by quadratic splines produces an  $f_0$  curve which, when concatenated, generates an intonation contour which is practically identical to the  $f_0$  contour on utterances consisting of sonorant segments which are both continuous and smooth. It was also argued that the quadratic spline

## CHAPTER 9. STANDARD YORÙBÁ PROSODY MODEL DESIGN

functions used for speech synthesis can be defined by a sequence of target points corresponding to the significant changes of the  $f_0$  thereby permitting easier computational analysis, manipulation and synthesis. *Campione et al.* (1997) also argued that, stylisation by quadratic splines produces a curve which is closer to the original  $f_0$  curve and theoretically, therefore, should be a more representative model of the  $f_0$  curve. In addition, *Dutoit and Leich* (1994) have shown that parameters based on linear models introduce artefacts in the resulting synthetic speech. Linear interpolation is also known to introduce undesirable angles at the concatenation points on the resulting  $f_0$  contours.

The work of *Levitt and Rabiner* (1971) further supports the use of functions with higher degree than the linear equation. In an experimental study, Levitt and Rabiner employed averaging and polynomial fitting to model sentence  $f_0$  contour, comparing the accuracy of fit as a function of the degree of the polynomial. Although they did not relate their finding to the synthesis of  $f_0$  contour of utterances, they show that the higher degree polynomials produce the better fit. This may be because higher degree polynomials make it easier to locate the turning points and trends on an  $f_0$  curve.

In respect of tone languages, *Shen et al.* (1993) have shown that the  $f_0$  turning points are a perceptual cue to Mandarin tones 2 and 3. They also show that, while overall height of the  $f_0$  curve may contribute to the distinctive phonetic characteristics of these two tones, the timing of the  $f_0$  turning points as well as the difference in frequencies between the peaks and valleys of an  $f_0$  curve contribute significantly to the accurate perception of tones. The implication of this finding is that, in order to model the  $f_0$  curve, at least for tone language, the turning points are important not only as a categorisation feature but also as a cue for the accurate perception of the pitch associated with each tone. Linear functions are unable to capture these  $f_0$  transitions that are important in the perception of pitch.

What all these point to is that there is a clear distinction between what is mathematically plausible, what is acceptable in engineering terms and what is practical when it comes to modelling perceptually significant aspects of  $f_0$ . As correctly noted by *Pierrehumbert* (2000), it may be the case that what counts perceptually and phonologically as the pitch of a tone from a mathematical point of view, may not have

a well-defined  $f_0$  correlate. This may be partly because a speaker's use of pitch is usually attributed to the non-linguistic factors, such as the personality of the speaker, which are unquantifiable. The next section documents our experiments conducted into stylisation and standardisation of SY tones.

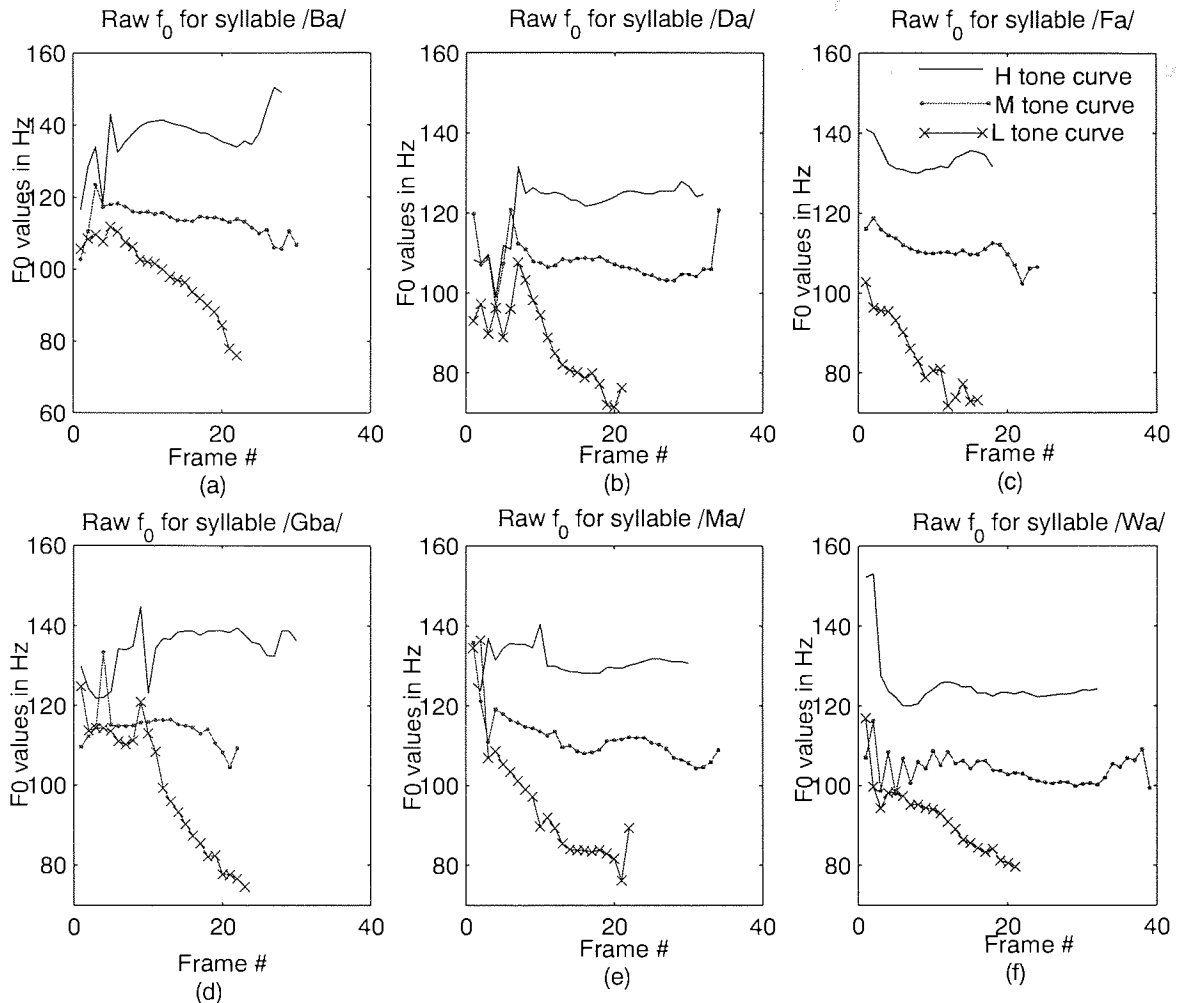
## 9.4 SY $f_0$ stylisation

We have chosen to implement  $f_0$  stylisation from the first principles. By this we mean that the  $f_0$  data corresponding to the tone associated with syllables are directly approximated by using various interpolation polynomials. Re-synthesised speech using these approximated  $f_0$  curves is then subjected to perceptual evaluation to determine the polynomial that produced the best speech quality while, at the same time, facilitating a transparent representation of the  $f_0$  curve. The pattern of the stylised curves will then inform the shape of the standardised symbol that will be used in representing the tonal inventories.

To achieve this, we first conducted an informal experiment in which we investigated the effects of the  $f_0$  curve on the perception of tone by manipulating the parameters of  $f_0$  curve on isolated syllables. The parameters manipulated include the turning points and the height of the  $f_0$  curve. In the experiment, the waveforms of pre-recorded isolated syllables were loaded into *Praat* (Boersma and Weenink, 2004) and their  $f_0$  curves were manipulated. The plots of the raw  $f_0$  curves of some SY CV type syllables, where  $V = a$ , are shown in Figures 9.8. A *MATLAB 6.5* program was written to read each file and interpolate the data using the *polyfit* command. Different interpolation functions were generated and experiments were performed to determine the best polynomial that describes the  $f_0$  data. The procedure for the interpolation experiment is discussed in the following subsections.

Informal listening tests were conducted to determine the effects of manipulating the peaks and valleys of the  $f_0$  curve on the perception of the syllable tone. The results of the experiment suggest that both the range and the turning points of  $f_0$  curves affect the perception of each of the three SY tones. For example, the H and L tones are clearly distinguished by the peak and valley on their  $f_0$  curve. The valley of the H tone occurs



Figure 9.8: Raw  $f_0$  of  $/CV/$  syllable where  $V=a$ 

early in the  $f_0$  and the peak occurs later. In the case of the L tone, the peak of the  $f_0$  curve occurs early and the valley occurs later. However, we seek to put our informal findings on a firm scientific footing by conducting a detailed experiment with the goal of determining the best  $n^{th}$  degree polynomial which, when used to approximate the  $f_0$  curve corresponding to SY tones, will preserve the perceptual features while at the same time facilitate computational manipulations and efficient synthesis.

### 9.4.1 Linear interpolation

In the first set of experiments, a linear function of the form  $y = a_1x + a_0$  is fitted into the  $f_0$  curve of the syllables. The parameter  $a_0$  is approximately equal to the mean of

the raw  $f_0$  value and it corresponds to the intercept on the frequency axis (in Hertz). The parameter  $a_1$  determines the slope of the  $f_0$  line. Figure 9.9a and Figure 9.10a depict the plots resulting from the linear interpolation into the  $f_0$  curves of the CV type syllables /*gba*/ and /*wa*/.

The parameters of the linear function interpolated into the raw  $f_0$  curve of the syllables are shown in Tables 9.2 and 9.1 respectively. Generally, the H tones have a positive slope while M and L tones have negative slopes with the slope of M being much smaller in magnitude than that of L. Moreover, the mean value of the  $f_0$  of the H tone is higher than that of M and L, with M having the smallest mean  $f_0$ .

Table 9.1: Coefficient for linear interpolation of  $f_0$  curve (Hz) /*Wa*/

Coefficient	Tones		
	<i>H</i>	<i>M</i>	<i>L</i>
$a_0$	134.41	108.79	124.01
$a_1$	0.25	-0.07	-2.04

Table 9.2: Coefficient for linear interpolation of  $f_0$  curve (Hz) /*Gba*/

Coefficient	Tones		
	<i>H</i>	<i>M</i>	<i>L</i>
$a_0$	128.74	118.70	126.46
$a_1$	0.376	-0.37	-2.28

Table 9.3: Coefficient for quadratic interpolation of  $f_0$  curve (Hz) for /*Wa*/

Coefficient	Tones		
	<i>H</i>	<i>M</i>	<i>L</i>
$a_0$	125.09	105.46	119.72
$a_1$	2.25	0.62	-0.92
$a_2$	-0.07	-0.03	-0.05

Table 9.4: Coefficient for quadratic interpolation of  $f_0$  curve (Hz) for /Gba/

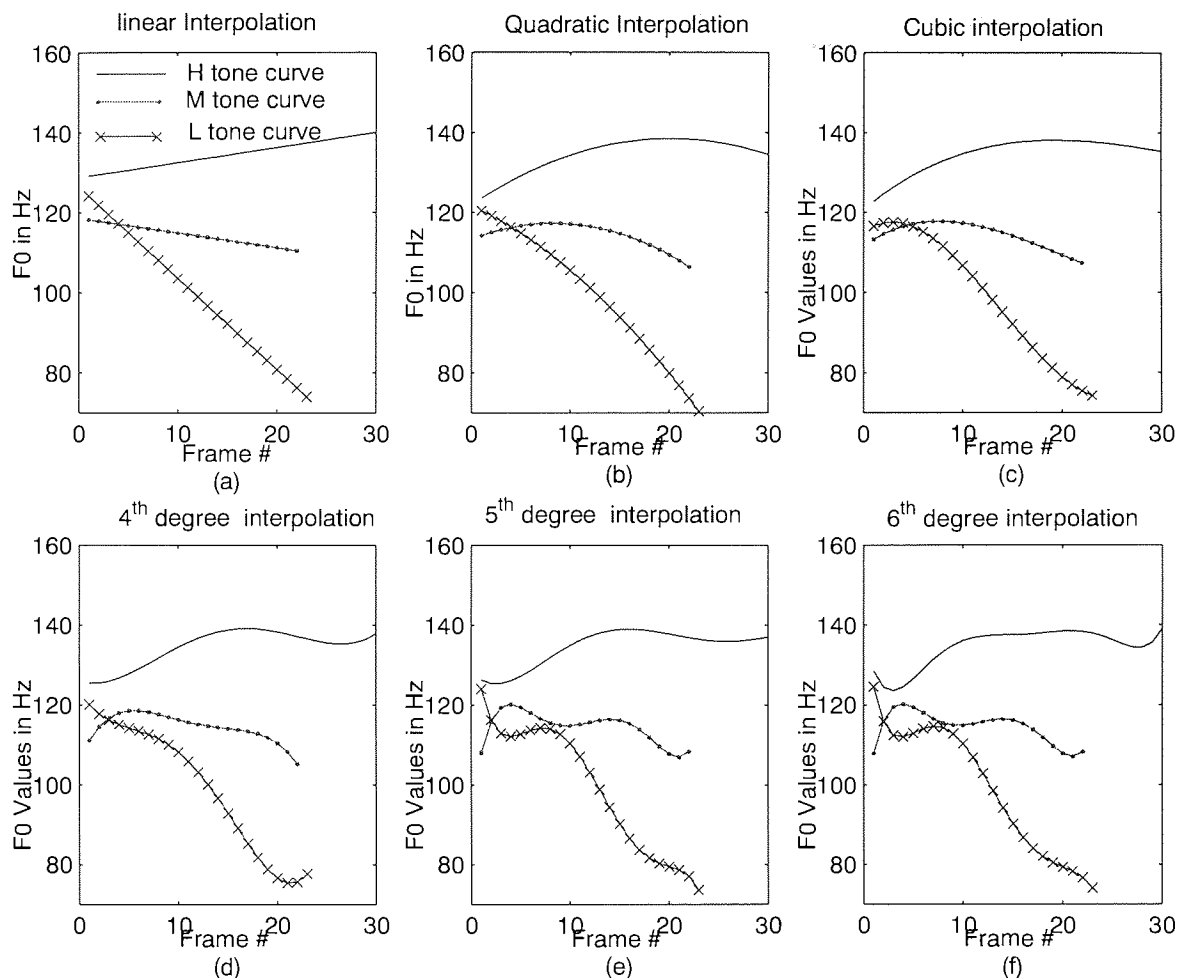
Coefficient	Tones		
	<i>H</i>	<i>M</i>	<i>L</i>
$a_0$	122.05	113.40	121.70
$a_1$	1.63	0.95	-1.14
$a_2$	0.04	-0.06	-2.05

### 9.4.2 Quadratic interpolation

For the quadratic interpolation experiment, a quadratic function of the form:  $y = a_2x^2 + a_1x + a_0x$  is interpolated into the raw  $f_0$  data of syllables carrying the H, M and L tones. Figures 9.9(b) & 9.10(b) depict the plot resulting from this interpolation. The parameter  $a_0$  is approximately equal to the mean of the raw  $f_0$  value where  $a_1$  and  $a_2$  determine the shape of the curves. The values of  $a_0$ ,  $a_1$  and  $a_2$  for the three tones associated with syllables /Wa/ and /Gba/ is shown in Table 9.3 and Table 9.4 respectively.

The mean value of the *H*, *M* and *L* tones represented as  $a_0$  are 122.05, 113.41 and 121.70 respectively. These results are consistent with what is obtained for the linear interpolation in the sense that the mean  $f_0$  value of the H tone is higher than that of the L tone whereas that of the L tone is higher than the M tone. Another distinguishing feature of the three tones is that both  $a_1$  and  $a_2$  are positive for the H tone, whereas they are negative for L tone. For the M tone, however,  $a_1$  is positive while  $a_2$  is negative. The  $f_0$  curve characteristics differ slightly for the syllable /Wa/.

As shown in Table 9.3 the mean values of the H, M, and L tones are 125.09, 105.46 and 119.72 respectively. These  $a_0$  values are relatively consistent with the mean value obtained for the syllable /Gba/ (see Table 9.3) apart from the fact that the  $a_1$  value of H tone, which is negative for /Wa/ syllable. The excursion, in terms of sign and magnitude, of the  $f_0$  curve for /Wa/ and /Gba/ are relatively similar but sufficiently distinct.

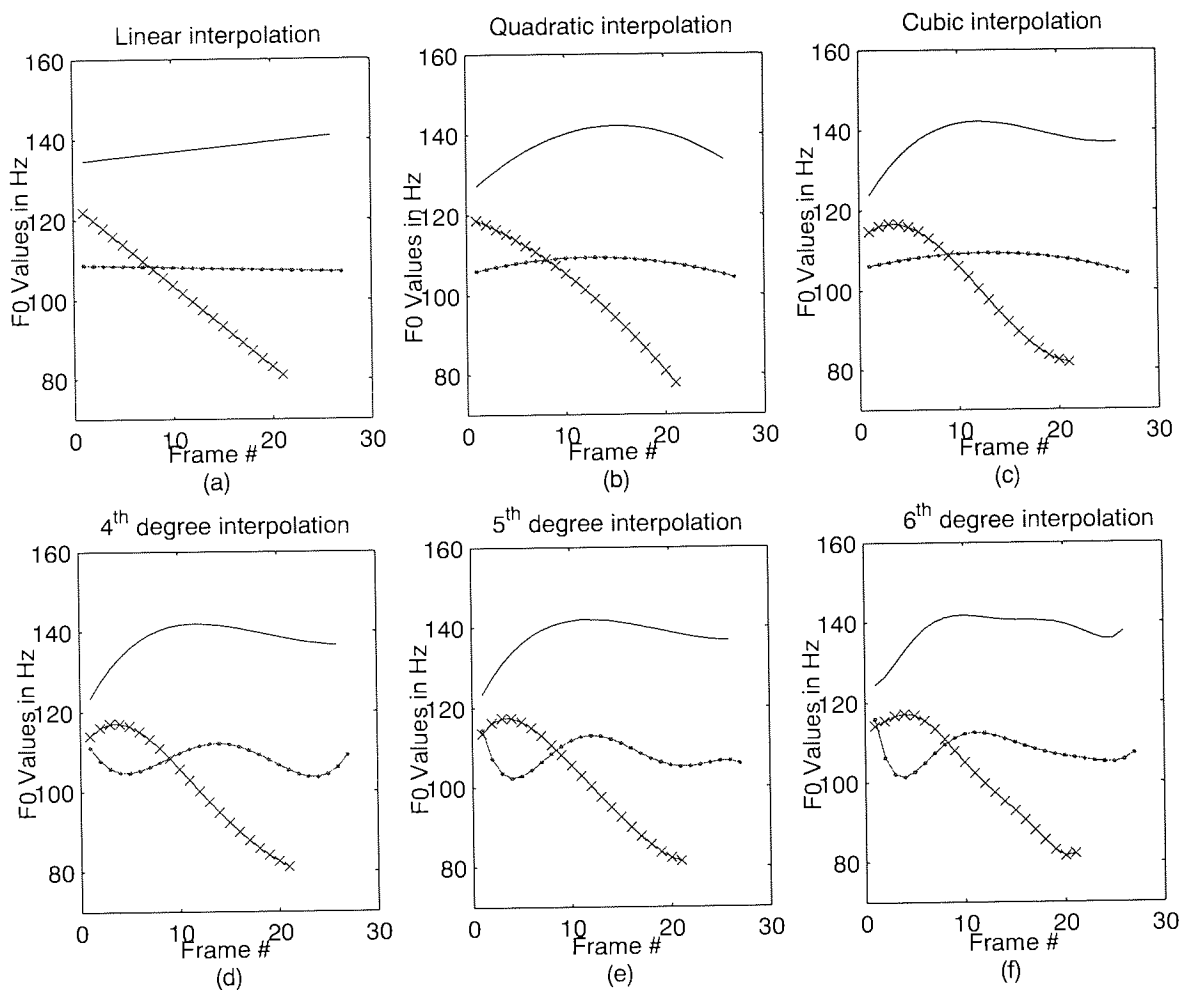
Figure 9.9: Interpolation into  $f_0$  curve of /Gba/ syllable

### 9.4.3 Higher degree interpolations

Further experiments were conducted using higher degree interpolation polynomials of the form:

$$y = \sum_{i=0}^{i=n} a_i x^i \quad (9.3)$$

where  $n > 2$  is the degree of the polynomial. We experimented with the 3<sup>rd</sup>, 4<sup>th</sup>, 5<sup>th</sup> and 6<sup>th</sup> degree interpolation polynomials. The resulting curves are shown in Figures 9.9c through 9.9f for syllable /Gba/ and Figures 9.10c through 9.10f for syllable /wa/. For the 3<sup>rd</sup> degree interpolation polynomials, that is  $y = a_3 x^3 + a_2 x^2 + a_1 x + a_0$ , the plot shows a clear and distinct pattern for each of the tones.

Figure 9.10: Interpolation into  $f_0$  curve of /*Wa*/ syllable

As shown in Figure 9.9c and Figure 9.10c, the H tone rises from about 120 *Hz* to 145 *Hz* and then flattens before falling slightly. The L tone on the other hand, peaks at around 118 *Hz* and falls steadily to reach a minimum value of about 75 *Hz*. The M tone is relatively stable, forming a concave pattern which peaks at about 110 *Hz* and about 115 *Hz* at the edges. The parameters of the polynomial are consistent with those obtained with the linear and quadratic interpolations.

Similar results were obtained for the 4<sup>th</sup> degree interpolation (see Figure 9.9d and Figures 9.10d). However, in this case, the  $f_0$  curve interpolated is rather unstable at the edges. This is particularly apparent in the M and L tone  $f_0$  curves. For the fifth and sixth degree interpolation polynomials, the higher order coefficients  $a_5$  and  $a_6$ , assumed zero or near zero values. This indicates that the factors they represent is not

contributing significantly to the shape of the  $f_0$  curve described by these polynomials. The seventh order interpolation does not produce stable polynomial parameters.

## 9.5 Evaluation of the stylisation functions

We have analysed and evaluated each of the above interpolation functions. The aim of the analysis was to determine the best  $f_0$  stylisation function that: (i) accurately approximated the  $f_0$  data, and (ii) adequately captures the perceptual quality in terms of producing a close-copy (Collier, 1990). In this way, we will be able to determine which of the functions best model the  $f_0$  curve of SY syllables while at the same time preserving their perceptual quality. Consideration is also given to the characteristics of the selected function as an effective mean for modelling utterances of larger linguistic units, such as multi-syllable words, phrases and sentences.

We applied two evaluations methods: (i) quantitative and (ii) qualitative. The quantitative evaluation method was based on the computation of Root Mean Square Error (RMSE). The qualitative evaluation is divided into two: (i) intelligibility evaluation and (ii) naturalness evaluation. The transcription error rate was used for intelligibility evaluation. For the naturalness evaluation, we used the mean opinion score (MOS) for evaluating the overall naturalness quality of the approximated versus natural  $f_0$  curves. Four subjects, all native speakers of SY, participated in the evaluation. The results of the evaluation are discussed in the following subsections.

Table 9.5: RMSE (Hz) result for each interpolation for syllable /Gba/

Tone	Linear	Quadratic	Cubic	4 <sup>th</sup> order	5 <sup>th</sup> order	6 <sup>th</sup> order	7 <sup>th</sup> order
H	0.89	0.74	0.73	0.70	0.69	0.66	Unstable
M	0.98	0.88	0.87	0.84	0.74	0.74	Unstable
L	0.99	0.91	0.82	0.73	0.55	0.54	Unstable

Table 9.6: RMSE (Hz) result for each interpolation for syllable /*Wa*/

Tone	Linear	Quadratic	Cubic	4 <sup>th</sup> order	5 <sup>th</sup> order	6 <sup>th</sup> order	7 <sup>th</sup> order
H	0.83	0.39	0.22	0.22	0.23	0.16	Unstable
M	0.69	0.64	0.64	0.43	0.27	0.20	Unstable
L	0.63	0.52	0.28	0.27	0.26	0.24	Unstable

### 9.5.1 Quantitative evaluation

The RMSE has the advantage of more accurately reflecting the performance of the stylisation function with respect to the actual value of the acoustic signal. The RMSE of the natural (raw) versus approximated  $f_0$  curves is computed using Equation 9.4:

$$RMSE = \sqrt{\frac{\sum_i^N (f_{0i}^{raw} - f_{0i}^{approx})^2}{N}} \quad (9.4)$$

where  $f_{0i}^{raw}$  and  $f_{0i}^{approx}$  are the raw and approximated  $f_0$  values respectively, and  $N$  is the number of  $f_0$  sample points. The results suggest that higher degree polynomials have lower RMSE than lower degree polynomials (cf. Tables 9.5 & 9.6). For example, while the RMSE for the linear interpolation is 0.89, that of quadratic is 0.74. The RMSE value continues to fall as the degree of the polynomial increases until the lowest value of 0.66 which is obtained for the 6<sup>th</sup> degree interpolation polynomial. This suggests that the higher the degree of the interpolation polynomial, the better the interpolation accuracy.

An observation of the plotted  $f_0$  curves in Figures 9.9(d-f) and 9.10(d-f) shows that interpolation polynomials with degree higher than three produces unstable behaviour at the  $f_0$  curve boundaries. For example, the plot for the 4<sup>th</sup> degree polynomials for the M tone shows a concave pattern at the beginning and toward the end of the  $f_0$  plot (see Figures 9.9(d-f) and 9.10(d-f)). The 5<sup>th</sup> and 6<sup>th</sup> degree polynomials produced more unstable patterns at the boundaries than the 4<sup>th</sup> degree polynomial. This unstable behaviour has a potential to introduce undesirable  $f_0$  patterns at the concatenation points on the resulting  $f_0$  contours.

In addition, the approximated  $f_0$  curves of the H tone using the 4<sup>th</sup> and 5<sup>th</sup> degree polynomials are similar to that of the 3<sup>rd</sup> degree polynomial. This suggests that the

higher degree polynomials are not adding more information to the represented curves than the 3<sup>rd</sup> degree. Hence, the parameters for the higher degree polynomials are redundant.

From a purely mathematical point of view, the smaller the error in the interpolation polynomial, the better the result. However, the results and observations from our experiments suggest that functions with a degree higher than 3 are not suitable due to their inherent instability and their redundant parameters. It is particularly important that the modelled  $f_0$  curve for syllables are stable at the boundaries. That is because when concatenating syllables to form larger linguistic units such as words, the boundaries play a crucial role in smooth transition of the acoustic waveform.

### 9.5.2 Qualitative evaluation

Based on the results of the quantitative evaluation discussed above, we reach a temporary conclusion that the linear, quadratic and cubic degree polynomials are the most appropriate candidates for the stylisation of SY tone curves. To select the best amongst these approximation polynomials, listening tests were conducted to determine the quality of speech produced by each  $f_0$  approximation function. The aim of the qualitative evaluation is to obtain native speakers' overall impression of the speech sound generated based on the approximate  $f_0$  curve when compared to natural speech.

Seven adult native speakers of SY participated in the qualitative evaluation. To ascertain their hearing ability before the experiment, some recorded natural speech sounds were played to them and they were asked to write down what they heard. Those who were unable to produce 100% accuracy in this test were excluded from the experiment. In the end, four adult male native speakers of SY participated in the qualitative test.

#### Intelligibility

For the intelligibility test, the speech sound with approximate  $f_0$  contour was played to the listeners. After a speech sound was played, the listener was asked to write down what he heard. In the intelligibility evaluation, the participants were not only required to identify the tones on synthesised syllables, they were also required to accurately



identify the syllables associated with each tone. The intelligibility score is computed using the following formula:

$$Intelligibility = \frac{T_{All} - T_{Wrong}}{T_{All}} \times 5.0 \quad (9.5)$$

where  $T_{All}$  is the total number of syllables presented for a particular tone and  $T_{Wrong}$  is the number of syllables that had been wrongly identified. The computation is used to determine the transcription error. The value 0 will be obtained from the computation when a respondent transcribes all the syllables in the synthesised utterance wrongly. The value 5 is obtained when all the syllables in the utterance are correctly transcribed.

The results for linear, quadratic and cubic interpolations are as recorded in Table 9.7. The average intelligibility score for linear, quadratic and cubic stylisation functions are 4.16, 4.42 and 4.58 respectively. The intelligibility score for natural speech is 4.99. That result shows that intelligibility of higher degree polynomials is slightly better than those of lower degree. It also shows that synthetic mid tone is better perceived than low tone, and low tone is better perceived than high tone.

Table 9.7: Intelligibility of SY syllables with natural and synthetic  $f_0$  curves

Syllable	stylisation function			
	Natural	Linear	Quadratic	Cubic
H tone	4.98	4.10	4.31	4.51
M tone	5.00	4.20	4.53	4.62
L tone	4.99	4.20	4.41	4.61
Mean	4.99	4.16	4.42	4.58

### Naturalness

The main purpose of the naturalness test is to determine how pleasant the speech is or, put in another way, how good the generated sound is in comparison with human speech. In the Mean Opinion Score (MOS) evaluation, listeners are asked to rate the quality of speech sound after the raw  $f_0$  curve has been replaced by the approximated ones. The listeners are presented with the speech samples of the raw and modified  $f_0$  curve in a random manner. They are then asked to rate their qualities on a scale of 1 to 5 (1 for very poor and 5 for excellent).

To conduct the test, the natural syllables (i.e. syllable with raw  $f_0$ ) and synthetic syllables were played at random. The MOS is the average of the opinion of all the listeners forming the mean of their impression about the naturalness of the speech sound.

The results are summarised in Figure 9.11. For example, for the H tone, the MOS of 3.25, 3.80 and 4.21 was obtained for linear, quadratic and cubic interpolation, respectively. The MOS is 4.98 for natural speech. The same pattern of results are repeated for low and mid tones. These results suggest that the acceptability of the synthetic  $f_0$  curve is far above average when the stylisation function is cubic.

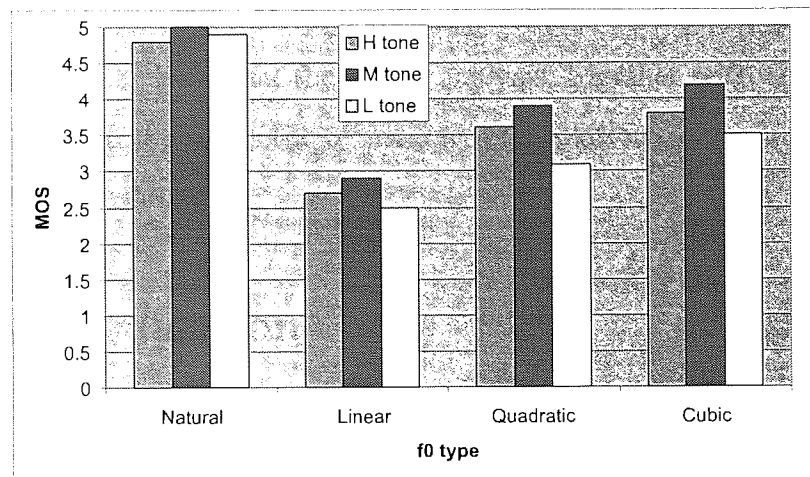


Figure 9.11: Mean opinion score for naturalness

### 9.5.3 Summary of stylisation experiment

The linear and quadratic functions provide a trivial mathematical representation of the  $f_0$  curves on SY tones. Basically they give a general course of the  $f_0$  curve transition which may be useful in tone recognition. And as observed by *Lu et al.* (2001), the requirement for intonation modelling is more demanding than in the case of tone recognition. An alternative is to use a spline based on linear or quadratic function. This option will further complicate the stylisation process and make it more difficult to manipulate the functions when processing the  $f_0$  curves for intonation synthesis. Moreover, when compared with the cubic function, the linear and quadratic functions produce less accurate  $f_0$  model.

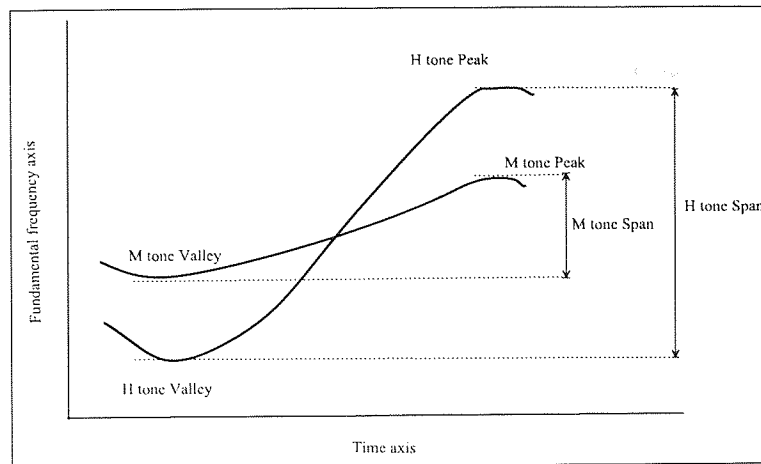
In addition, the cubic function is able to model the two turning points which are essential to the perception of the H and L tones. These two points correspond to the peak and valley on the  $f_0$  curves. We also observed, during the experiment, that when the 3<sup>rd</sup> degree interpolation function for the M tone is shaped so that there is an early valley and late peak, SY listeners still perceived the mid pitch accurately. The accuracy of the perception remains unchanged when compared to the original  $f_0$  curves as long as the peak and valley are within the  $f_0$  range of the M tone. This observation, at least for the M tone, confirms the findings by *Connell et al.* (1983) that the recognition of tone appears to be stable over a wide range of pitch changes. In this regard, we can say that, overall, the shape of the M tone generally follows that of the H tone, although at a much lower scale. Based on the above analysis, together with the results of the qualitative evaluation, we select the 3<sup>rd</sup> degree interpolation polynomial for modelling the  $f_0$  curve of SY syllables.

## 9.6 Standardisation

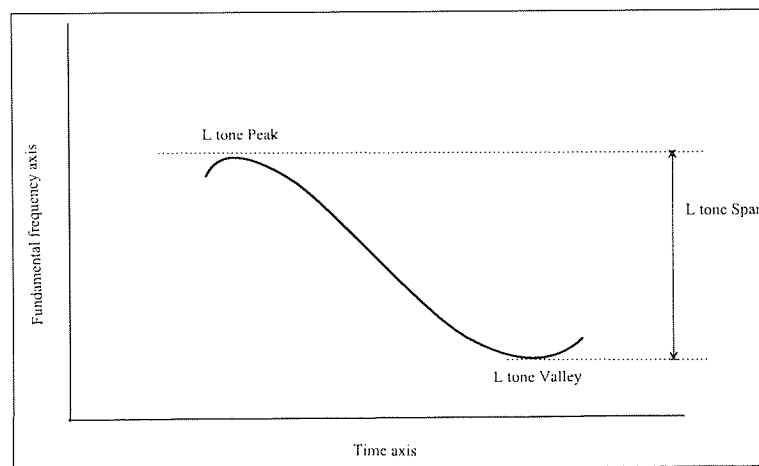
The various pitch movements that occur in our stylisation experiment may differ from each other in terms of the excursions of the  $f_0$  curves (*Connell et al.*, 1983). In order to achieve a certain degree of generalisation, it is necessary that the different types of tone movements should be given standard specifications of their relevant features (*Collier*, 1990).

We observed that the  $f_0$  curve of the High (H) and Low (L) tones are distinguished by two turning points, i.e. the *valley*, which is the turning point at the lowest  $f_0$  value on the curve, and the *peak*, which is the turning point at the highest  $f_0$  value on the curve. The Mid (M) tone exhibits different behaviour with respect to the turning points. Generally, there are four important features that characterise the  $f_0$  curve of a tone. These include:

1. the turning points on the  $f_0$  curve,
2. the spatial structure of the  $f_0$  curve,
3. the range of the  $f_0$  values, i.e. the difference between the minimum and maximum  $f_0$  values, and



(a) Abstract curves of H and M tones



(b) Abstract curves of L tones

Figure 9.12: Abstract curve of the three Yorùbá tones

4. the span of the  $f_0$  curve, i.e. the difference between the point on the time axis, when the minimum and maximum  $f_0$  values occur.

Based on the above analysis, we can represent each of the three SY tones as depicted in Figure 9.12. The most important aspect of this representation is the location of the peak and valley on the  $f_0$  as well as their range. In the H and M tones, the valleys occur before the peaks (cf. 9.12a). However, the range (the absolute magnitude of the difference between the peak and valley) of the H tone is larger than that of an M tone by an order as much as 2. In the case of an L tone (cf. 9.12b), the peak occurs before the valley and the  $f_0$  range is approximately equal to that of H tone.

We need symbolic parameters for each type of SY tones for the standardisation. These parameters are intended to be a generalisation of the  $f_0$  curves into the symbolic

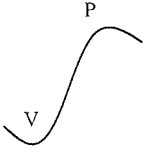
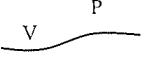
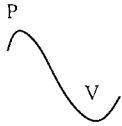
Tone	Standardised curve	Peak-Valley sequence
H		VP
M		VP
L		PV

Figure 9.13: Signature for the standardised tones

categories which facilitate the application of phonological rules for describing the pattern of sequence of such symbols. Based on the above stylisation experiment, therefore, we formulate the signature shown in Figure 9.13 for the standardisation of the three SY tones.

## 9.7 Abstract intonation pattern generation

The purpose of the stylisation and standardisation processes is to represent each tone symbolically. The symbolic representation allows us to apply phonological rules to predict the pattern and spatial structure corresponding to a sequence of tones.

At the phonological level, it is possible to describe an utterance intonation in terms of interaction between tones, such as High (H), Mid (M), and Low (L), which are discrete entities. Perceptually, however, tones are mainly distinguished by relative changes in pitch which is signalled by the position of the turning points on the  $f_0$  curve, as well as by the relative height of the frequency contour (*Connell et al.*, 1983). We assume that these turning points and relative  $f_0$  contour heights can be manipulated computationally in order to model the spatial structure of an intonation pattern.

A number of works have documented the heuristics which describe tone interaction in SY (*Bámgbósé*, 1966; *Hombert*, 1976, 1977; *Connell and Ladd*, 1990). For example, *Courtenay* (1971) suggests that the tonal system of SY can be clarified considerably if

Table 9.8: Terracing rule in SY (*Courtenay* (1971))

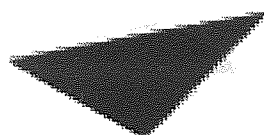
Rule	Context
Tone $\Rightarrow$ n	after pause
H $\Rightarrow$ +6	L <sub>-</sub>
+3	M <sub>-</sub>
0	L <sub>-</sub>
L $\Rightarrow$ -8	H <sub>-</sub>
-4	M <sub>-</sub>
0	L <sub>-</sub>
M $\Rightarrow$ +3	L <sub>-</sub>
+3	H <sub>-</sub>
0	M <sub>-</sub>

it is recognised that the terracing of both the high and mid tones is one process. Based on this assumption, she proposed the set of phonological rules in Table 9.8 for tone interaction in SY. The numerical values in the rules indicate the degree to which the pitch of the target tone is increased or decreased. For example, the rule  $H \Rightarrow 6 / L$  states that the pitch of an H tone is increased to level 6 if it precedes an L tone.

In a phonetic study of SY tones on VCV nouns, *Hombert* (1976) documented the approximate shape of the  $f_0$  curves for six tone patterns as shown in Figure 9.14. These findings show that the tonal signature of an  $f_0$  curve can be linked using phonological rules and it can be used to predict the intonation pattern of the utterance corresponding to a sequence of tones. Consequently, an algorithm can be derived to generate the intonation pattern using the standardised  $f_0$  curve of tones discussed above. The peaks and valleys of the sequence of such tones will be the basic input to the algorithm. We proposed an algorithm, formulated around the notion of a Skeletal Tree (S-Tree), which continuously determines the highest peaks and lowest valleys in the intonation pattern and assigns them to nodes in the S-Tree.

Based on the various phonological rules and heuristics described above, we use the following template for specifying the rules:  $A \Rightarrow B/C$ . This rule template is interpreted as follows: tone  $A$  is realised as tone  $B$  in the context  $C$ . For example, the rule  $T1 \Rightarrow T3/T2_{-}$  means that the tone  $T1$  is realised as tone  $T3$  if it is preceded by tone  $T2$ .  $T2_{-}$  is the context and the underscore indicates the position occupied by the tone on the left hand side of the rule.

Using the rule template, we specify SY phonological tone rules as shown in Table



**Aston University**

Illustration removed for copyright restrictions

Figure 9.14: Graphical representation of co-articulated tones according to *Hombert* (1976)

9.9. Rule set (i) in Table 9.9 can be interpreted as follows: the  $f_0$  curve of any tone that is not preceded by a tone is unchanged. By implication, the first tone in an utterance or an isolated syllable retains its tonal property. Rule 3 in rule set (iii) i.e.  $L \Rightarrow Low(L)/L_-$  is interpreted as follows: the pitch of a syllable carrying a low tone will be further lowered if it is preceded by another low tone carrying syllable.

The rules in Table 9.10 are an extension of the rules in Table 9.9. They implement and illustrate important intonation phenomena such as *H rising* and *L lowering* in SY intonation. This shows that the rules in Table 9.9 can be nested and used to generate the abstract pattern and structure of the intonation contour corresponding to a sequence of tones.

### 9.7.1 Skeletal tree generation algorithm

After describing the  $f_0$  curve of each syllable in terms of peak and valley, the next step is to design an algorithm for generating the Skeletal Tree (S-Tree). Our algorithm generates S-Tree using the rules shown in Table 9.9. For example, using rule 3 in Rule Set (iii), the highest peak in a sequence of four low tone syllables, i.e.  $L_1L_2L_2L_4$ , will be associated with the first syllable. Subsequent syllables in the sequence will have

Table 9.9: Phonological rules for SY tone interaction

Rule set	Rule No.	Rule Specification
<i>i</i>	1	$H \Rightarrow H/\Phi H$
	2	$M \Rightarrow M/\Phi M$
	3	$L \Rightarrow L/\Phi L$
<i>ii</i>	1	$H \Rightarrow High(H)/L_$
	2	$H \Rightarrow SlightlyHigh(H)/M_$
	3	$H \Rightarrow Low(H)/H_$
<i>iii</i>	1	$L \Rightarrow VeryLow(L)/H_$
	2	$L \Rightarrow SlightlyHigh(L)/M_$
	3	$L \Rightarrow Low(L)/L_$
<i>iv</i>	1	$M \Rightarrow SlightlyHigh(M)/L_$
	2	$M \Rightarrow SlightlyLow(M)/H_$
	3	$M \Rightarrow M/M_$

where  $\Phi$  denotes a blank.

lower peaks, with the fourth having the lowest. Similarly, the lowest valley will be associated with the tone of the fourth syllable and the valley of all preceding tones will be higher, with the first syllable having the highest valley. By applying this process recursively, peaks and valleys (corresponding to the turning points on waveforms) are computed and assigned to the nodes on the resulting S-Tree.

An S-Tree is generated in such a way that nodes of the tree that represent valleys on the waveform also contain a series of smaller subtrees. Therefore, a node on the tree is itself a root of a subtree which describes a smaller portion, or subwaveform, of

Table 9.10: Depiction of intonation phenomenon in SY tone interaction

phonological Rule	Intonation phenomena
$H \Rightarrow SlightlyLow(H)/H_$	H Downdrift
$M \Rightarrow SlightlyLow(M)/M_$	M Downdrift
$L \Rightarrow SlightlyLow(L)/L_$	L Downdrift
$H \Rightarrow VeryHigh(H)/L_$	H raising
$M \Rightarrow High(M)/L_$	Extension of H raising to M
$H \Rightarrow H/M_$	H raising in the context of M
$H \Rightarrow Low(H)/HL_$	Downstep
$L \Rightarrow VeryLow(L)/H_H$	L lowering
$L \Rightarrow Low(L)/M_M$	L lowering in the context of M



the intonation waveform. For each subwaveform, a root (i.e. the lowest valley) and its two child nodes (highest left and right peaks) are determined and assigned to be the left and right nodes of the tree respectively. To maintain an order of magnitude, the greater peak is assigned to the left node and the smaller peak is assigned to the right nodes. This process continues until all valleys have been processed and the terminal nodes on the tree (i.e. the leaves) are all peaks. The following assumptions are made in deriving the S-Tree generation algorithm.

- Each tone has exactly one peak and one valley on its stylised  $f_0$  curve.
- Any two peaks in an intonation waveform must have a valley in between them.
- The peaks and the valleys correspond to the turning points on the intonation waveform.

The main pseudocode for the S-Tree generation algorithm is shown in Figure 9.15. The subroutines for computing the lowest (or deepest) valleys and highest peaks are shown in Figures 9.16 and 9.17 respectively. The following subsection illustrates how the algorithm is used to generate S-Tree for a typical SY single-phrase sentence.

## 9.8 Skeletal tree generation for SY sentences

The S-Tree generation algorithm is demonstrated using the Yorùbá sentence, “Wón s̀ì tún p̀òwè ìbàdàn.” (meaning “[*They also said proverb Ibadan.*] *They also said Ibadan proverb*”). This sentence contains five words. The first three words have one syllable each. The fourth and fifth words contain two and three syllables respectively. This makes a total of 8 syllables in the utterance. The tones on the syllables are (HLHLMLLL).

The transcript of the S-Tree generation process for the sentence is shown in Figure 9.18. In the first step of the algorithm, the tone of each syllable is associated with a peak (see Figure 9.18a). Adjacent peaks are then paired from left to right and a valley is associated with each pair. The lowest valley and highest peaks are then determined and assigned to nodes on the tree in a recursive manner. At the end of the algorithm, an S-Tree that depicts the shape and structure of the intonation contour is generated (cf. Figure 9.18b).

```

Begin
  Initialisation
    n := number of syllables in the sentence;
    For j := 1 to n
       $P_j$  := Peak of Syllable j;
       $V_j$  := Valley of Syllable j;
      Create a peak-valley (PV) structure by associating a valley with
      each peak pair;
      if (j > 1) Then
         $(PV)_{j-1} := [(P_{j-1}, P_j), V_{j-1}]$ ;
      Endfor
    Head := 1; Tail := n;
    Do
    {
      Do
      {
        Start := Head;
        End := Tail;
        Pos := Find-Deepest-Valley(Head, Tail);
        CurrentValleyPosition := Pos;
        (Note: If Head = 1 and Tail = n,  $V_{Pos}$  is the root node)
        Partition the peak-valley structure into two at Pos;
         $P_L$  := Find-Highest-Peak(Start, Pos);
        Head := Position( $P_L$ );
         $P_R$  := Find-Highest-Peak(Pos, End);
        Tail := Position( $P_R$ );
        If ( $P_L < P_R$ ) Then Swap( $P_L, P_R$ );
        Create a left node for  $V_j$  and label it as  $P_L$ ;
        Create a right node for  $V_j$  and label it as  $P_R$ ;
      } While (More peak-valley structure to the left of  $V_{Pos}$ )
    } While (More peak-valley structure to the right of  $V_{Pos}$ )
    End
  End

```

Figure 9.15: Main pseudocode for S-Tree generation

```

Function Find-Deepest-Valley(DHead, DTail)
Begin
  DeepestValleyPos := 0;
  Found := false;
  k=DHead;
  While(!Found AND k<DTail)
  {
    If ( $P_k \in ((P_k, P_{k+1}), V_k)$ ) is associated with last L tone
      Then Found := True; Return k;
    If ( $P_k \in ((P_k, P_{k+1}), V_k)$ ) is associated with last M tone
      Then Found := True; Return k;
    If ( $P_k \in ((P_k, P_{k+1}), V_k)$ ) is associated with last H tone
      Then Found := True; Return k;
    k = k + 1;
  }
End Find-Deepest-Valley

```

Figure 9.16: Pseudocode for finding deepest valley

```

Function Find-Highest-Peak(DHead, DTail)
Begin
  HighestPeakPos := 0;
  Found := false;
  k=DHead;
  While(!Found AND k<DTail)
  {
    If ( $P_k \in ((P_k, P_{k+1}), V_k)$ ) is associated with first H tone
      Then Found := True; Return k;
    If ( $P_k \in ((P_k, P_{k+1}), V_k)$ ) is associated with first M tone
      Then Found := True; Return k;
    If ( $P_k \in ((P_k, P_{k+1}), V_k)$ ) is associated with first L tone
      Then Found := True; Return k;
    k = k + 1;
  }
End Find-Highest-Peak

```

Figure 9.17: Pseudocode for finding highest peak

As depicted in the S-Tree, the deepest valley on the intonation waveform for the sentence is associated with the 7<sup>th</sup> syllable (i.e. bà) in the sentence. This valley is bounded by peak  $P_1$  (associated with first H tone in the sentence) to the left and peak  $P_8$  (associated with last L tone in the sentence) to the right. Peaks  $P_1$  and  $P_8$  dominate<sup>1</sup> the other peaks on the left and right of the intonation waveform respectively. On the

<sup>1</sup>i.e. higher peak from where lower peaks are referenced.

portion of the waveform dominated by peak  $P_1$ , the lowest valley is  $V_6$ . This valley is bounded by  $P_1$  and  $P_7$  to the left and right respectively. Subsequent valleys on the peak dominated by  $P_1$ , in order of magnitude, are  $V_5$ ,  $V_4$ ,  $V_3$ ,  $V_2$  and  $V_1$ .  $V_5$  is bounded by peaks  $P_1$  and  $P_6$ , while  $V_1$  is bounded by peaks  $P_1$  and  $P_2$ . Note that peak  $P_8$  is a terminal peak.

The S-Tree generated at the end of this process represents the peaks and valleys on the intonation waveform as well as the self-embedding structure of the waveform. This tree depicts the abstract structure of the intonation contour shown in Figure 9.18c. Note that, peaks on the S-Tree are arranged in a relative order of magnitude and their subscripts indicate their position on the waveform. For example, peak  $P_1$  occurs before  $P_2$ ,  $P_2$  occurs before  $P_3$ , etc.

## 9.9 Extension of the S-Tree algorithm

The S-Tree generation algorithm illustrated above can easily model the spatial structure of the intonation waveform for sentences containing one phrase. For sentences with more than one phrase, the algorithm must be extended. One way to do this is to construct an S-Tree for each phrase in a multi-phrase sentence and then combine the individual S-Trees to produce the complete S-Tree for the whole sentence. Sentences with more than one phrase will contain more than one intonation phrase, each with its own  $f_0$  peaks and valleys. Given that two S-Trees,  $ST_1$  and  $ST_2$ , are to be combined to form another S-Tree  $ST_3$ , we need to determine the location of the highest peak and lowest valley on  $ST_3$  with respect to peaks and valleys on  $ST_1$  and  $ST_2$ .

In a recent work, *Smith* (2004) has demonstrated that the intonation and timing patterns of *non-sentence-initial* phrases are moderated by sentence initial phrases. Based on this view, we can speculate the behaviour of the  $f_0$  pattern of multi-phrase sentences. In a two-phrase sentence, for example, we can assume that the highest peak on intonation contour of the first phrase in the sentence is relatively higher than that of the second phrase. Furthermore, the deepest valley of the final phrase in a sentence will be deeper than that of the initial phrase.

To generate the R-Tree for a multi-phrase sentence, we use a composite approach.

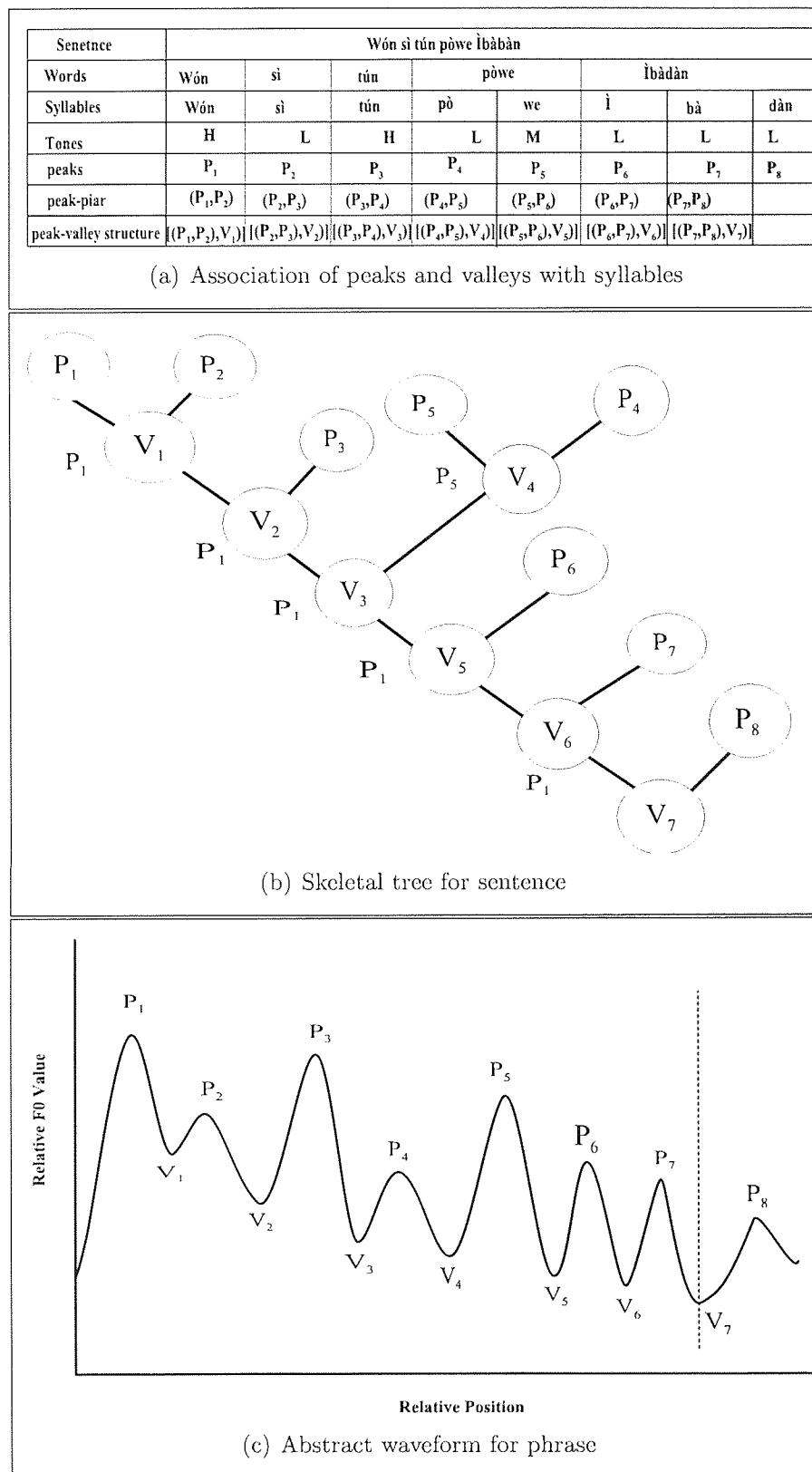
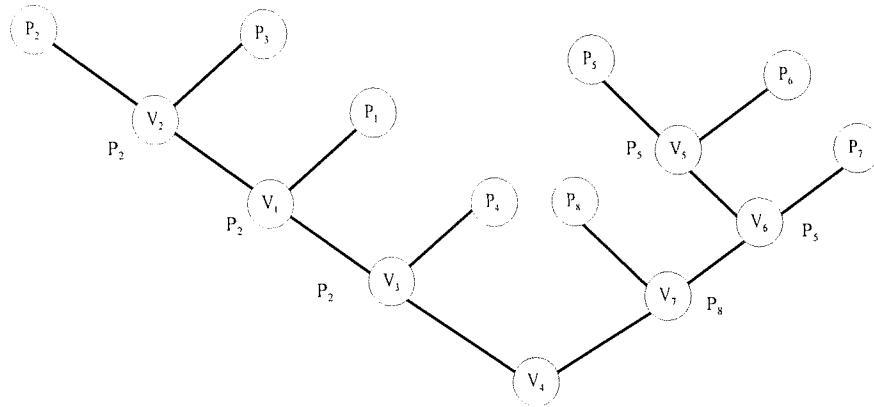


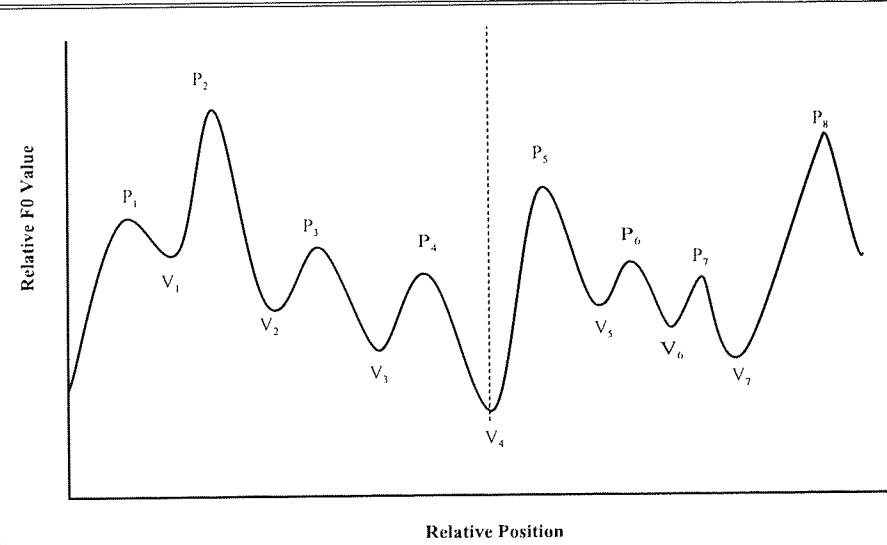
Figure 9.18: Transcript of S-Tree generation for the SY sentence “Wón sì tún pòwe Ìbàdàn.”

Phrase	Bàbá àgbè tita kòkò													
Words	Bàbá		àgbè		tita		kòkò							
Syllables	Bà	bá	à	gbè	tí	ta	kò	kó						
Tones	L	H	L	L	M	M	L	H						
peaks	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>	P <sub>6</sub>	P <sub>7</sub>	P <sub>8</sub>						
peak-piar	(P <sub>1</sub> ,P <sub>2</sub> )	(P <sub>2</sub> ,P <sub>3</sub> )	(P <sub>3</sub> ,P <sub>4</sub> )	(P <sub>4</sub> ,P <sub>5</sub> )	(P <sub>5</sub> ,P <sub>6</sub> )	(P <sub>6</sub> ,P <sub>7</sub> )	(P <sub>7</sub> ,P <sub>8</sub> )							
peak-valley structure	[(P <sub>1</sub> ,P <sub>2</sub> ),V <sub>1</sub> ]		[(P <sub>2</sub> ,P <sub>3</sub> ),V <sub>2</sub> ]		[(P <sub>3</sub> ,P <sub>4</sub> ),V <sub>3</sub> ]		[(P <sub>4</sub> ,P <sub>5</sub> ),V <sub>4</sub> ]		[(P <sub>5</sub> ,P <sub>6</sub> ),V <sub>5</sub> ]		[(P <sub>6</sub> ,P <sub>7</sub> ),V <sub>6</sub> ]		[(P <sub>7</sub> ,P <sub>8</sub> ),V <sub>7</sub> ]	

(a) Association of peaks and valleys with syllables



(b) Skeletal tree for phrase "Bàbá àgbè tita kòkò"

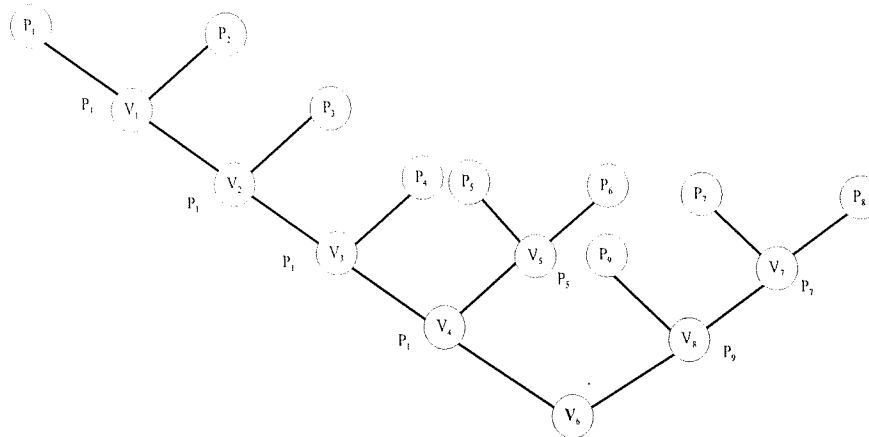


(c) Abstract waveform for phrase "Bàbá àgbè tita kòkò"

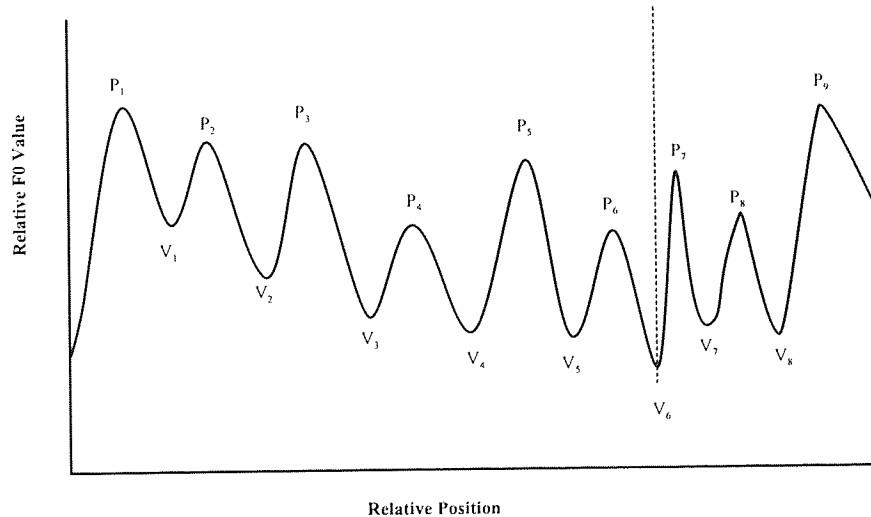
Figure 9.19: Transcript of S-Tree generation for phrase "Bàbá àgbè tita kòkò"

Phrase	kiótó m̀pé kòkò tíwón.									
Words	kiótó			m̀pé		kòkò		tíwón.		
Syllables	ki	ó	tó	m̀	pé	kò	kò	ti	M	wón
Tones	H	H	H	L	H	L	H	M	H	
peaks	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>	P <sub>6</sub>	P <sub>7</sub>	P <sub>8</sub>	P <sub>9</sub>	
peak-pair	(P <sub>1</sub> ,P <sub>2</sub> )	(P <sub>2</sub> ,P <sub>3</sub> )	(P <sub>3</sub> ,P <sub>4</sub> )	(P <sub>4</sub> ,P <sub>5</sub> )	(P <sub>5</sub> ,P <sub>6</sub> )	(P <sub>6</sub> ,P <sub>7</sub> )	(P <sub>7</sub> ,P <sub>8</sub> )	(P <sub>8</sub> ,P <sub>9</sub> )	(P <sub>9</sub> ,P <sub>9</sub> )	
peak-valley structure	(P <sub>1</sub> ,P <sub>2</sub> ,V <sub>1</sub> )	(P <sub>2</sub> ,P <sub>3</sub> ,V <sub>2</sub> )	(P <sub>3</sub> ,P <sub>4</sub> ,V <sub>3</sub> )	(P <sub>4</sub> ,P <sub>5</sub> ,V <sub>4</sub> )	(P <sub>5</sub> ,P <sub>6</sub> ,V <sub>5</sub> )	(P <sub>6</sub> ,P <sub>7</sub> ,V <sub>6</sub> )	(P <sub>7</sub> ,P <sub>8</sub> ,V <sub>7</sub> )	(P <sub>8</sub> ,P <sub>9</sub> ,V <sub>8</sub> )		

(a) Association of peaks and valleys with syllables



(b) Skeletal tree for phrase "ki ótó m̀pé kòkò tíwón"



(c) Abstract waveform for phrase "ki ótó m̀pé kòkò tíwón"

Figure 9.20: Transcript of S-Tree generation for phrase "ki ótó m̀pé kòkò tíwón"

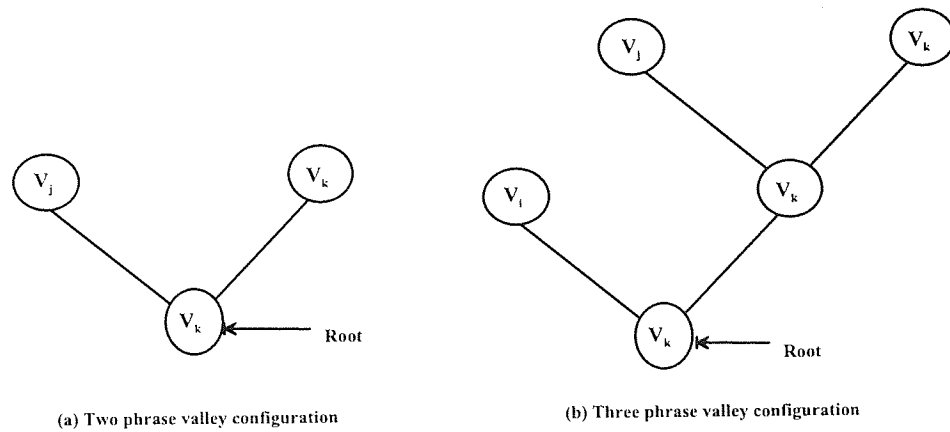


Figure 9.21: Relational organisation of valleys in multi-phrase utterance

Given a sentence made up of two phrases  $Pr_1$  and  $Pr_2$  and  $Pr_1$  is the sentence initial phrase and  $Pr_2$  is the sentence final phrase. If we apply the idea that the  $f_0$  contour generally follow a downtrend pattern as suggested by *Viana et al.* (2003), we can assume that the highest peak in a two-phrase utterance will be located in the first phrase, i.e.  $Pr_1$ , while the lowest valley will be located in the second phrase, i.e.  $Pr_2$ . If  $V_j$  and  $V_k$  are the deepest valleys in  $Pr_1$  and  $Pr_2$  respectively, then  $V_k$  will be at a lower level than  $V_j$ .

In the S-Tree generation algorithm, the underlying principle is that lower valleys appear towards the right node while higher valleys appear to the left of the root of the tree. Going by this principle, we can represent the two valleys  $V_j$  and  $V_k$  using the configuration in Figure 9.21a. This configuration shows that the valley  $V_j$  is dominating  $V_k$  thus making  $V_k$  the root of the new S-Tree. The S-Tree for  $Pr_1$  and  $Pr_2$  can then take their root from the vertex  $V_j$  and  $V_k$  respectively.

We illustrate the above algorithm with the sentence, ““Bàbá àgbè tita kòkó, kí ótó m̀òpé kòkó tí wón”” (meaning “[*Father farmer has sold cocoa, before he knows that cocoa expensive*] *The farmer has sold cocoa, before knowing that the price of cocoa has increase*”). The sentence comprises two phases separated by a comma. The S-Tree generation transcript for the two phrases and their abstract waveforms are shown in Figure 9.19 and Figure 9.20 respectively. The deepest valley in the intonation waveform, for the combined S-Tree, falls on the second phrase, i.e. “*kí ótó m̀òpé kòkó tí wón*”. This valley is identified as  $V_6$  in the second phrase. If we index the phrases



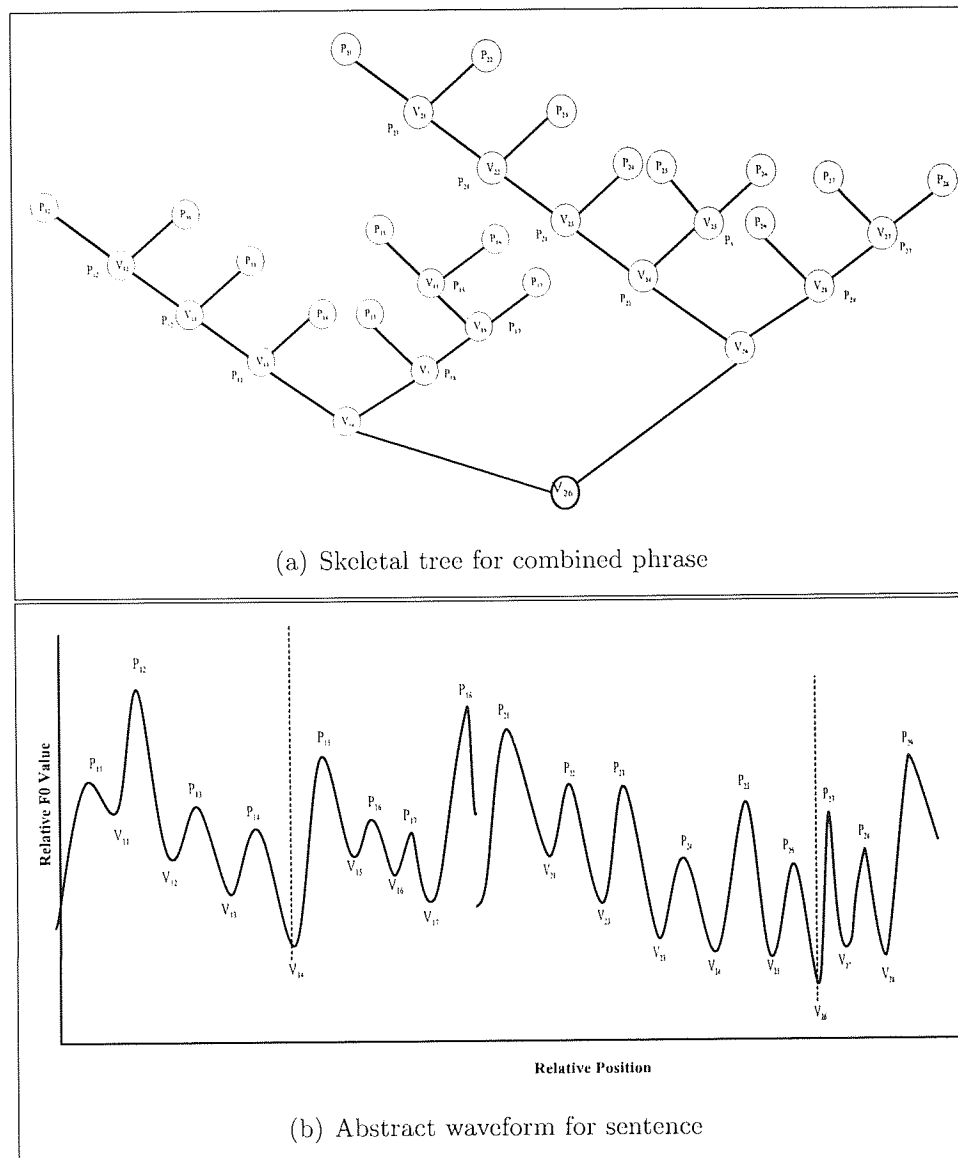


Figure 9.22: Abstract intonation waveform for sentence “*Bàbá àgbè tita kòkó, kí ótó m̀òpé kòkó ti wón*”

in the order of their occurrence in the sentence, i.e. the first phrase has index 1 and second phrase has index 2, then we can re-number the peaks and valleys using a two-dimensional index  $(p, r)$ . Here  $p$  is the phrase index and  $r$  is the peak or valley index. Using this scheme, the deepest valley on the intonation contour, which is located at the sixth valley on the second phrase, will be labelled  $V_{(2,6)}$ . The combined tree and abstract waveform for the two phrases are shown in Figure 9.22a and Figure 9.22b respectively.

The above technique can be applied to a three-phrase utterance  $Pr_1$ ,  $Pr_2$ , and  $Pr_3$  in which case the lowest valley will be associated with  $Pr_3$  and the tree will be

configured as shown in Figure 9.21b. In this case,  $Pr_1$  is the sentence initial phrase,  $Pr_2$  is the sentence medial phrase and  $Pr_3$  is the sentence final phrase. This technique can be applied recursively on multi-phrase utterance to generate the S-Tree.

This model assumes that look-ahead and preplanning strategies are employed in the generation of the abstract intonation contour. Formally specified, this means that, the computation of the parameter of an intonation object  $Y_i$  depends on the ‘accessible’ properties of its predecessor, i.e.  $Y_{i-1}$ , if such is available. Thus the computation of the  $f_0$  curve of a syllable depends on the curve of the preceding syllable, that of a phrase depends on the preceding phrase.

All these features must be explained either by features spreading at the phonological level, by computation of some parameters of a higher level constituent, by interpolation between one target and the next, or by temporal overlap in the realisation of segments of the utterance’s intonation. The hierarchical representation allows a more direct interaction between sub-constituent and superordinate structures, where larger scale components determine the scope for the smaller scale components and where the implementation of tonal events is performed on the basis of upcoming, as well as preceding, events and is sensitive to syntactic and semantic structuring.

If enough parameters are introduced in the formula for pitch accent computation in the tonal sequence model, it will be possible to handle both the phrase sequence position and the difference between different boundary conditions. Consecutive lowering of individual phrases could be handled by a rule which downsteps the first pitch in each component relative to the first pitch on the preceding one, or, if it turns out that only three textual positions are distinguished, the first pitch in the initial and final phrase could be marked for “initial” and “final”, respectively, and initial pitch accent in the medial components can be given the same “medial” value.

## 9.10 Features of the S-Tree model

We identify the following as the unique attributes of the S-Tree presented in this work.

- It provides an elegant conceptual scheme that allows for the abstract representation of local and global attributes of  $f_0$  contour in relative terms. For example, the  $f_0$  peak at the rightmost leaf of the tree indicates the relative point where

the *highest* global  $f_0$  value occurs and the valley at the root is the point where the *lowest*  $f_0$  value occurs. This relation is not linear though. The term *highest* and *lowest* are relatively specified in the description. These two points give a clue to the relative global range of the  $f_0$  contour. Also the peaks and valleys of individual tones are easily distinguishable in the S-Tree. This makes it easy to locate and analyse specific domain of intonation phenomena, both local and global, and relate them to the  $f_0$  contour.

- It facilitates the abstract visualisation of intonation contour as a multi-dimensional model in which different dimensions are represented and can be analysed independently. This is because, although the S-Tree is generated using the tonal attributes of the sentence, it also acts as a structure on which intonation and related prosody attributes such as duration and intensity dimensions can be anchored. This makes it possible to add information to the S-Tree and modify its structure at various stages of prosody generation for an utterance. Therefore, information in respect of different dimensions of speech prosody can be added whenever they become available.

This attribute of the model facilitates the integration of identified utterance specific attributes, parameters, and functions into a speech waveform in an iterative manner. A subtree of the S-Tree headed by any node contains all the intonational attributes logically associated with that node. By traversing the subtree headed at any node, we can update and access all the relational values, i.e.  $f_0$ , duration and intensity, logically associated with that node and relate them to corresponding locations on the waveform. This makes it easier to observe how each variable affects the quality of synthetic speech prosody.

- Most phonological theories use a tree data structure to represent phenomena in spoken and written languages. In this respect, therefore, S-Tree can provide a systematic way of relating textual information to the phonological structure and then to the prosody of synthetic speech. This is because it provides a hierarchically-structured and functionally-organised model in which elements can be related to linguistic entities and hence be interpreted linguistically. The operators made available in tree algebra can easily be extended to transform tree representation to a waveform and *vice versa*. This feature of S-Trees facilitates a transparent and computationally effective implementation of the model as well as an intuitive interface showing the relationship among the prosodic entities.
- The incorporation of stylised  $f_0$  curve as the primitive on which S-Tree generation is based makes it possible to account for perceptually relevant  $f_0$  curve in the intonation contour.

## 9.11 Summary

In this section, we have discussed a crucial part in our R-Tree based approach to prosody modelling. Specifically, we have presented the process for stylising and stan-

standardising SY tone from first principles. We have also presented an algorithm that takes the parameters of standardised sequences of tones, i.e. peaks and valleys, and generates the Skeletal Tree (S-Tree). The next step is to reduce the abstract intonation pattern represented by the S-Tree into its phonetic equivalent. In order to achieve this, we need to determine the exact numerical values of the  $f_0$  peaks and valleys as well as their position in the duration axis on the generated S-Tree. This is the aim of the subsequent chapters in this path.

# Chapter 10

## Intonation modelling

In the second stage of the *R-Tree* based prosody modelling, we compute the numerical values of perceptually significant transition on the selected dimension. While the *S-Tree* generation algorithm is linguistically-driven, the computation of the numerical values for the intonation dimension is generic, in that it can be data-driven or rule-driven. The computed value for each peak and valley of the intonation contour is then incorporated into the corresponding peak ( $P_i$ ) or valley ( $V_i$ ) on the *S-Tree*. The actual intonation contour is obtained by joining the computed points using interpolation. The method for computing the points on the *S-Tree* is able to take advantage of the abstract computation that has already been carried out in the previous step. In the case of intonation modelling, the *S-Tree* can be viewed as a model which organises the intonation units (i.e. tones) into a coherent structure. This structure forms the basis for realising the actual intonation waveform.

We compute the numerical values for the intonation dimension using a fuzzy logic based modelling technique. The application of fuzzy logic to speech signal modelling is not novel in itself. In speech recognition, for instance, a number of works (*Pal and Majumder, 1977; De Mori et al., 1979; Demichelis et al., 1983; O'Brien, 1993; Kosanović et al., 1996; Breining, 2001*) have shown that fuzzy logic models are capable of modelling the acoustic and perceptual aspects of speech signal.

In speech synthesis, *Raptis and Carayannis (1997)* demonstrated the application of a fuzzy logic rule-based approach to formant speech synthesis. Also, *Jitca et al. (2002)* have shown that an improved speech synthesis can be obtained by applying a fuzzy

## CHAPTER 10. INTONATION MODELLING

logic based method to the computation of the first two formants (i.e.  $F1$  and  $F2$ ) of phonemes. Recently, *Lin et al.* (2003) combined a fuzzy logic based approach with a recurrent neural network for modelling Mandarin Chinese prosody. All these studies showed that the application of fuzzy logic to speech signal produced more robust models and models that are easy to implement, upgrade and interpret.

Our approach differs from all of the above mentioned fuzzy logic based speech signal modelling in two respects. Firstly, we used fuzzy logic as a technique for realising the intonation contour in the context of concatenative speech synthesis. Secondly, we have combined the *R-Tree* technique with fuzzy logic into a unified framework. Within this framework, the fuzzy logic transforms the linguistic description of intonation phenomena into numerical values for realising the intonation contour of an utterance. The features of the  $f_0$  curve of each syllable in an utterance are related by fuzzy rules which operate on numerically rendered acoustic cues. This enables our model to account for the intonation phenomena synthesised by the  $f_0$  contour, with a degree of evidence varying along a continuum.

### 10.1 The training data

For the purpose of developing our fuzzy logic based intonation model, we selected a set of data from the speech database discussed in Chapter 8. The selected data comprises 460 isolated SY syllables (230 each for two adult native male speakers of SY). The  $f_0$  values of each syllable are extracted and a third degree polynomial  $P_3(f_0)$  was interpolated into the  $f_0$  values. The canonical peak and valley of each syllable, which correspond to the maximum and minimum turning points of  $P_3(f_0)$  respectively, is then extracted. To obtain the data for the natural realised  $f_0$ , a total of 30 *statement* sentences were selected from the database. The syllables for each sentence are drawn from the 230 syllables mentioned above. Using the *Praat* (*Boersma and Weenink, 2004*) software, the location of the peak and valley of the  $f_0$  curve of each syllable is obtained from the  $f_0$  contour of each sentence.

## 10.2 Fuzzy model design

There are two stages in developing a fuzzy-logic control model: *structure identification* and *parameter identification* (Takagi and Sugeno, 1985). In *structure identification*, the input and output of the system as well as their interconnections are determined. During this process, all the measurable input variables that are known to contribute to the value of each output variable are determined. This is normally done using a divide-and-conquer strategy whereby an input is selected from all the inputs and its influence on the target output variable is determined. During the parameter identification stage, the optimal consequence parameters are determined. This is normally formulated as a linear programming problem which can be solved by the least squares method. We discuss the structure and parameter identification of our model in the next sub-sections.

### 10.2.1 Related models

In the structure identification stage, we first examined the structure of intonation models put forward by experts on Yorùbá (Ladd, 1987; Connell and Ladd, 1990; Harrison, 2000; Lánúran and Clements, 2003) and Mandarin Chinese (Shih, 2000). These models shed light on three aspects of intonation generation: (i) the input variables of an intonation model, (ii) interactions and relationships between these variables during the intonation generation process and, (iii) how the resultant effects of the interactions between these variables culminate to produce the intonation contour corresponding to a perceived intonation phenomena.

#### Categorical and gradient models

Lánúran and Clements (2003) described two theoretical models that can be used for modelling the interaction between downstep and H tone rising, which are the central intonation phenomena in SY speech. These are the “Categorical” and “Gradient” models. Structurally, the two models are similar in that they account for similar input and output variables. They both use the  $f_0$  values of tones and the positions of syllables in the sentence to determine the amount of downstep that the  $f_0$  contour will undergo. Using the theoretical description of the categorical model provided in Lánúran and

*Clements* (2003) and *Stewart* (1993), we can express the amount of downstep that the  $m^{\text{th}}$  downstepped H tone,  $H_m$ , will undergo as:

$$H_m = f_D(H_{m-1}; m) \quad (10.1)$$

where  $m$  is an integer whose value equals to the number of downsteps undergone by the  $m^{\text{th}}$  downstepped H tone (i.e.  $H_m$ ).  $f_D$  is a function that relates the  $f_0$  of the previous downstepped H tone to  $m$ . This model is not suitable for our purpose because it does not provide an interpretation of the integer  $m$  in terms of the actual  $f_0$  contour.

Unlike the categorical model, the gradient model defines downstepping pattern as a gradual decay towards an abstract reference line, or asymptote (*Lieberman and Pierrehumbert*, 1984). In this approach, the value of any H tone in a downstepping sequence,  $H_m$ , is given by the equation:

$$H_m = d \times (H_{m-1} - r) + r \quad (10.2)$$

where  $H_{m-1}$  is the  $f_0$  value of the preceding H tone,  $d \in [0, 1]$  is the downstep coefficient, and  $r$  is the value of the reference line towards which the  $f_0$  of the H tone declines. Equation 10.2 assigns  $f_0$  values to downstepping H tone from left to right within the downstep span. For example, consider the case in which the parameters  $d$  and  $r$  are set at 0.7 and 100Hz respectively. If the first H tone in the sequence  $H_1 L_1 H_2 L_2 H_3$ , i.e.  $H_1$ , has an  $f_0$  value of 150Hz then  $H_2$  will have the  $f_0$  value of 135Hz =  $(0.7 \times (150 - 100) + 100)$ Hz. The next downstepped H tone, i.e.  $H_3$ , will have the  $f_0$  value of 124.5Hz =  $(0.7 \times (135 - 100) + 100)$ Hz. The model describes an exponentially decaying  $f_0$  curve in which each downstep is proportionally identical to the preceding one in terms of distance from the reference line  $r$ . Another attribute of the model is that later downstep intervals are progressively smaller than earlier ones, and tend to become vanishingly small as the reference line is approached.

### Shih's model

*Ladd* (1990) has, however, observed that the categorical and gradient models are complementary in the sense that the gradient model can be regarded as providing speaker-specific  $f_0$  parameters that can be used in the interpretation of the integer sequence



provided by the categorical model. *Shih* (2000) has proposed a model which takes Ladd's observation into account. Shih's model is an exponentially decaying model for declination in Mandarin Chinese. The model was developed around the assumption that exponentially decaying declination can handle longer sentences than a time constant model based on fixed  $f_0$  values. On the basis of this assumption, *Shih* (2000) proposed the following equation for modelling downtrend phenomena in Mandarin Chinese:

$$P_i = \alpha P_{i-1} + \beta(P_1 - \alpha(P_1)) \quad (10.3)$$

where  $P_i$  is the  $f_0$  value of a given syllable and  $\beta$  is given by:

$$\beta = \frac{\mu}{(P_1 - \alpha(P_1))} \quad (10.4)$$

In this model the  $f_0$  value of a given syllable is estimated from the  $f_0$  value of the preceding tone (i.e.  $P_{i-1}$ ).  $\alpha$  and  $\mu$  are speaker-dependent parameters. Shih's model is able to account for the phenomena whereby the  $f_0$  values decline faster in the early section of an utterance, then asymptote to a value which is appropriate to the speakers  $f_0$  range. The parameter  $\alpha$  controls how fast the  $f_0$  value approaches the asymptote. The smaller the value of  $\alpha$ , the steeper the decline. The parameter  $\mu$  controls the asymptote value. For example, when  $\mu$  equals zero the  $f_0$  value asymptote to zero and when  $\mu$  equals the difference between the initial  $f_0$  value, i.e.  $P_1$ , and  $\alpha(P_1)$ , the asymptote value is  $P_1$ , with zero declination. An important attribute of Equation 10.3 is that  $P_1$ ,  $\alpha$ , and  $\beta$  are intuitively linked to the initial value, rate of declination, and the asymptote value of an intonation contour, respectively.

### 10.2.2 Model structure identification

The transition and spatial structure of the intonation contour of an SY utterance is a result of complex interactions between a number of intonation phenomena. When observed individually, the phenomena can be categorised into two kinds. The first kind are phenomena that cause the  $f_0$  contour to rise (we call these *R-phenomena*). This includes phenomena such as *H rising* and *f<sub>0</sub> resetting*. The second kind are phenomena that cause the  $f_0$  contour to fall or lower (we call these *L-phenomena*).

This includes phenomena such as *downstepping*, *final lowering* and *L lowering*. These phenomena exhibit complex interaction during speech production. Their individual contribution to the transition and spatial structure of the intonation waveform is very difficult to isolate and describe quantitatively. The resultant effect of these phenomena on the generated  $f_0$  contour can vary over a wide spectrum depending on the resultant strength of each kind of phenomena.

On one end of the spectrum is a situation where only *R-phenomena* are present and no *L-phenomena*. The resultant effects of this will move the trajectory of the intonation contour towards a higher  $f_0$  value. On the other end of the spectrum is a situation where the *L-phenomena* are acting exclusively. The resultant effect of this will move the intonation contour trajectory towards a lower  $f_0$  value. At the middle of the spectrum, the *R-phenomena* and *L-phenomena* will act simultaneously and with equal strength. In such a situation, the two kinds of phenomena will neutralise the effect of each other on the  $f_0$  contour and the aggregate effect will leave the  $f_0$  trajectory unperturbed. Other effects of the interactions between *R-phenomena* and *L-phenomena* on an  $f_0$  contour cannot be described in categorical terms. For example, if the *R-phenomena* has higher strength than the *L-phenomena*, the overall effect is for the  $f_0$  contour to be raised by an amount proportional to the difference in strengths. This proportion can only be described qualitatively using linguistic terms such as positive large, negative small, etc.

We observed in our data that the transition of the  $f_0$  contour depends on the combination of tones in an utterance and the minimum  $f_0$  value perceptually associated with a tone (henceforth  $f_0$  BASE). The tone perceived for any  $f_0$  value also depends on the  $f_0$  values of the preceding tone. We factored these observations into Shih's model (Shih, 2000) and the resulting model is depicted in Equation 10.5. Our model allows us to use the canonical peak/valley of the  $f_0$  curve of a tone to compute its realised  $f_0$  pattern in the intonation contour of an utterance with which it occurs. This accounts for transition from one tone pattern to another and makes it possible for the model to capture co-articulatory intonation phenomena such as the carry-over effects (Xu, 1999a, 1997).

$$f_{i+1}^R = \alpha \times \left( \frac{f_j^c}{f_{Last}^T} \right) \times (f_{Last}^T - BaseT) + BaseT \quad (10.5)$$

The parameter  $\alpha$  controls the rate at which the  $f_0$  contour asymptote to the minimum perceptible  $f_0$  value, i.e.  $BaseT$ , of a tone  $T$ . Our experiments show that  $\alpha$  is speaker dependent and  $|\alpha| \leq 1.0$ . The value of  $\alpha$  is further moderated by the ratio of the canonical peak of the  $f_0$  curve of the next tone to be concatenated, i.e.  $f_j^c$ , and the  $BaseT$  value. This ratio ensures that a proportionally smooth rate of decay is maintained throughout the course of the  $f_0$  contour transition.

A limitation of this equation is that, it does not modulate the asymptote value of  $f_0$  contour in the same manner as the rate of asymptote. Based on the assumption that each tone has its own span in the intonation space, we proposed that the  $f_0$  trajectory of each tone will asymptote to the  $BaseT$  of that tone. This implies that the  $f_0$  trajectory of tone H will asymptote to  $BaseH$ , that of M will asymptote to  $BaseM$  and that of L will asymptote to  $BaseL$ . The manner of approach for each tone's  $f_0$  towards the asymptote line is determined by the parameter  $\beta$ . Based on this view, we amend the model in Equation 10.6 to:

$$f_{i+1}^R = \alpha \times \left( \frac{f_j^c}{f_{Last}^T} \right) \times (f_{Last}^T - BaseT) + \beta \left( \frac{f_{Last}^T}{f_i^R} \right) + BaseT \quad (10.6)$$

The important variables and parameters in this new model are explained in Table 10.1.

Let:

$$R = \left( \frac{f_j^c}{f_{Last}^T} \right) \times (f_{Last}^T - BaseT), \quad S = \left( \frac{f_{Last}^T}{f_i^R} \right), \quad \text{and} \quad C = BaseT$$

then we can write Equation 10.6 as:

$$f_{i+1}^R = \alpha R + \beta S + C \quad (10.7)$$

This model accounts for the intonation phenomena by considering not only the tonal constraint but also the relative degree with which the constraint is exhibited based on the exact  $f_0$  value of the tone. When compared with the reviewed models, our model produced a better prediction of the intonation contour (see Figure 10.1). Moreover, our model is able to predict the  $f_0$  values for a sequence of tones irrespective

Table 10.1: A description of the fuzzy model variables and parameters in Equations 10.5 &amp; 10.6

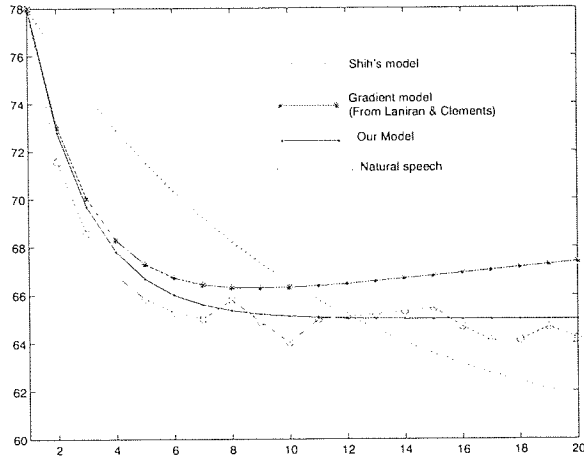
Symbol	Description
$f_j^c$	The canonical peak/valley of the $f_0$ curve of the next syllable to be concatenated (read from database)
$f_i^R$	The realised (computed) peak/valley of syllable in position $i$ of the intonation contour
$f_{i+1}^R$	The realised (computed) peak/valley of the next syllable in the intonation contour
$f_i^c$	The canonical peak/valley of the $f_0$ curve of the syllable in position $i$ (read from database)
$BaseT, T \in \{H, M, L\}$	The minimum value of $f_0$ peak that is perceptible as tone H, M, and L respectively
$f_{Last}^T$	The realised (computed) peak/valley of last the syllable with tone T
$\alpha$	Parameter that controls the rate at which the $f_0$ contour declines
$\beta$	Parameter that controls where the $f_0$ contour will asymptote to on the fundamental frequency dimension

of the tone combination or length of the utterance. For example, the result of  $f_0$  values computed for a sentence of twenty syllables with different tone combinations is shown in Figure 10.2.

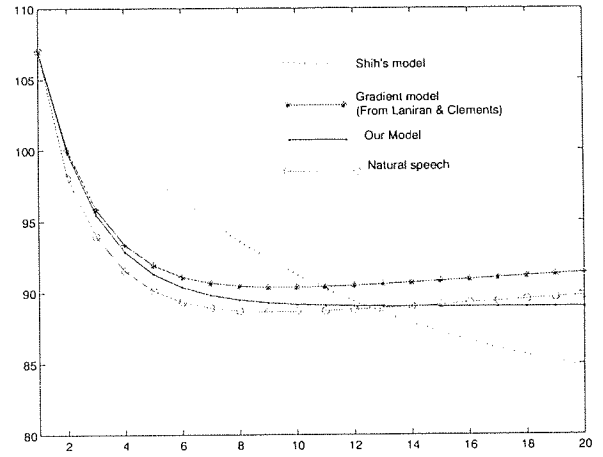
### Fuzzy model structure for SY intonation

Based on the result of the above analysis, the structure of our fuzzy model is made up of two premise variables (i.e. *TonCon* and *RelPos*) and one consequence variable (i.e.  $f_{i+1}^R$ ) which is the realised  $f_0$  value. The input variable *TonCon* is the difference between the canonical peak/valley of the  $f_0$  curve of the tone to be concatenated and the peak/valley of the last realised tone of the same type. The variable *RelPos* is the numerical value associated with the relative position of each syllable in a sentence.

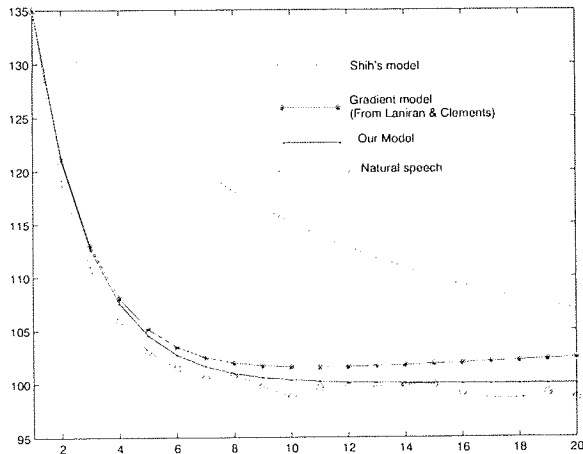
For the model to operate as a fuzzy model, each of the inputs (or premise variables) must take on qualitative values. Using MATLAB 6.5, for each sentence in our database, we conducted a graphical analysis of the change in  $f_0$  values and the position of syllable. We then partitioned the universe of discourse of the variable *TonCon* into five categories. The linguistic terms used to label each of the categories are: *Negative Large (NL)*, *Negative Small (NS)*, *No Change (NC)*, *Positive Small (PS)*, *Positive Large*



(a) L tone sequence



(b) M tone sequence



(c) H tone sequence

Figure 10.1: Natural and predicted  $f_0$  peak plots for intonation models for the same tone sequence

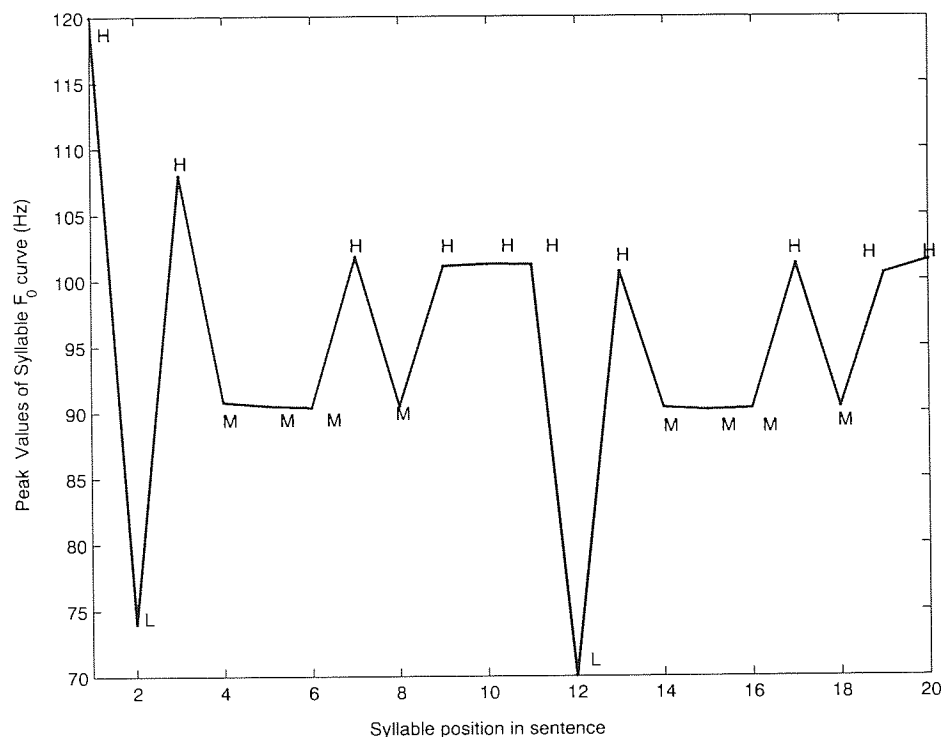


Figure 10.2: Computation of  $f_0$  peak on a sentence containing syllables with the H, M, and L tone sequence

$(PL)\}$ . We represent this fuzzy set as  $A_j = \{NL, NS, NC, PS, PL\}; j = 1, 2, 3, 4, 5$  (see Table 10.2).

Table 10.2: Linguistic label for *TonCon*

Description	Label	$A_j$
Negative Large	NL	$A_1$
Negative Small	NS	$A_2$
No Change	NC	$A_3$
Positive Small	PS	$A_4$
Positive Large	PL	$A_5$

*RelPos* is computed with respect to the beginning of a sentence. Let  $n$  be the length of a sentence, i.e. number of syllables in the sentence, to compute the *RelPos* for the syllable in position  $i$ , we use the function:

$$RelPos = \frac{i}{n} \quad (10.8)$$

Table 10.3: Linguistic label for *RelPos*

Description	Label	$B_k$
Near	N	$B_1$
Far	F	$B_2$

*RelPos* is a concordance index between the length of a sentence and the position of syllables in the sentence. We partitioned *RelPos* space into two subspaces: *Near* and *Far*. The first syllable in a sentence is the nearest syllable and we can assign 1.0 as the approximate value of its membership to the set *Near* and 0.1 to its membership in *Far*. Following the same procedure used for *TonCon*, the linguistic terms describing the *RelPos* premise variable are  $\{Near (N), Far (F)\}$ . We represent this fuzzy set as  $B_k = \{N, F\}; k = 1, 2$  (see Table 10.3).

We considered a number of membership functions such as triangular, Gaussian bell-shaped, clipper-parabola and the trapezoid (*Mitaim and Kosko, 2001*). The trapezoid membership function was selected due to its simplicity and the fact that it sufficiently describes the Universe of Discourse (UoD) for the input parameters *TonConV* and *RelPos*. We defined the trapezoidal function for the variable *TonCon* as a four-tuple (*Mitaim and Kosko, 2001*)  $(l_j, ml_j, mr_j, r_j)$  where  $ml_j \leq mr_j \in \mathbb{R}$ . The variables  $l_j > 0$  and  $r_j > 0$  denote the distance of the support of a function to the left and right of  $ml_j$  and  $mr_j$ , the centre of which is  $m_j = 1/2(ml_j + mr_j)$ . The degree to which a crisp *TonCon* value belongs to the fuzzy set  $A_j$ , i.e.  $\mu_{A_j}(TonCon) \in [0, 1]$ , is computed using the membership function:

$$\mu_{A_j}(x = TonCon) = \begin{cases} 1.0 - \frac{ml_j - x}{l_j} & \text{if } ml_j - l_j \leq x \leq ml_j \\ 1.0 & \text{if } ml_j \leq x \leq mr_j \\ 1.0 - \frac{x - mr_j}{r_j} & \text{if } mr_j < x \leq mr_j + r_j \\ 0.0 & \text{otherwise} \end{cases} \quad (10.9)$$

A separate set of the membership functions in Equation 11.4 is defined over the Universe of Discourse (UoD) of the *H*, *L* and *M* tones'  $f_0$  value space as shown in Figures 10.3(a), 10.3(b) and 10.3(c) respectively.

The membership function for the variable *RelPos*,  $\mu_{B_k}(RelPos)$ , is defined as:

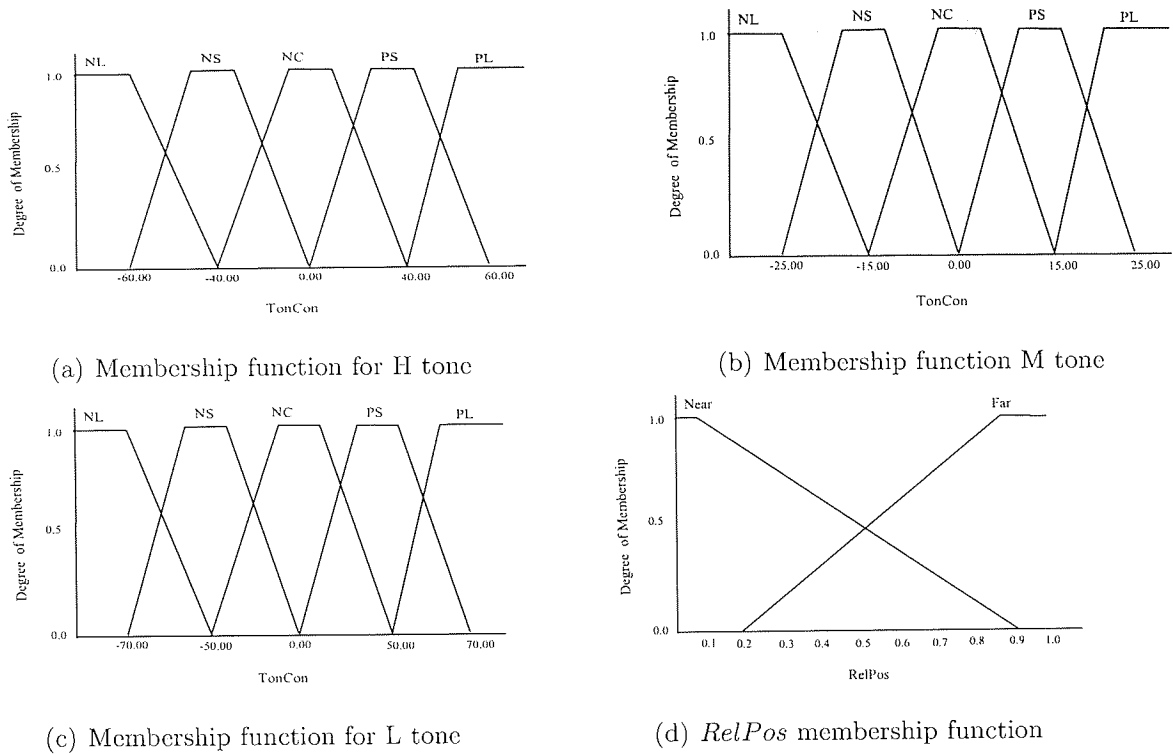


Figure 10.3: Membership functions for the premise variables

$$\mu_{B_1}(x = RelPos) = \begin{cases} 1.0 & \text{if } x < 0.2 \\ \frac{0.9-x}{0.9} & \text{if } 0.2 < x \leq 0.9 \\ 0.0 & \text{if } x > 0.9 \end{cases} \quad (10.10)$$

$$\mu_{B_2}(x = RelPos) = \begin{cases} 0.0 & \text{if } x < 0.2 \\ \frac{x-0.2}{x} & \text{if } 0.2 < x \leq 0.9 \\ 1.0 & \text{if } x > 0.9 \end{cases} \quad (10.11)$$

The model structure is specified as shown in Equation 10.12.

$$R_n : \text{ IF } (TonCon \text{ IS } A_j) \text{ AND } (RelPos \text{ IS } B_k) \\ \text{ THEN } f_{i+1}^n = \alpha R + \beta S + C \quad (10.12)$$

In Equation 10.12,  $R_n$  is the label for the  $n^{th}$  rule in the fuzzy rule base. *TonCon* and *RelPos* are the premise variables computed as described in Table 10.1.  $f_{i+1}^n$  is the output from the  $n^{th}$  implication and  $\alpha$  and  $\beta$  are the consequence parameters. The crisp output of the model (i.e the inferred  $f_0$  value for the peak/valley of the  $(i + 1)^{th}$



syllable in an utterance) is computed by taking the weighted average of the consequence of each rule as follows:

$$\bar{f}_{i+1} = \frac{\sum_{n=1}^N w_n \times f_{i+1}^n}{\sum_{n=1}^N w_n} \quad (10.13)$$

where  $N$  is the total number of rules in the fuzzy rule base and  $w_n$  is computed as:

$$w_n = \mu_{A_j}^n(TonCon) \wedge \mu_{B_k}^n(RelPos) \quad (10.14)$$

The fuzzy inference system discussed above has been proven to be a universal approximator which is capable of approximating any real continuous function to any degree of accuracy (Kosko, 1994; Weitian, 2001).

### Model parameter identification

The next step in our fuzzy intonation modelling is consequence *parameter identification*. The premise variable *TonCon* can assume five different linguistic values, (cf. Table 10.2) while the variable *RelPos* can assume two linguistic values, (cf. Table 10.3). This gives a total of  $5 \times 2 = 10$  possible combinations of premise situations as the input to the fuzzy model. The parameters of each rule (i.e.  $\alpha$  and  $\beta$ ) are the only distinguishing factors. The input-output relation is then formulated as:

$$f_{i+1} = \frac{\sum_{n=1}^N (A_j^n(TonCon) \wedge B_k^n(Relpo) * (\alpha R + \beta S + C))}{\sum_{n=1}^N A_j^n(TonCon) \wedge B_k^n(Relpo)} \quad (10.15)$$

Let:

$$\gamma_n = \frac{(A_j^n(TonCon) \wedge B_k^n(Relpo))}{\sum_{n=1}^N A_j^n(TonCon) \wedge B_k^n(Relpo)} \quad (10.16)$$

we can write Equation 10.15 as:

$$\begin{aligned} f_{i+1} &= \sum_{n=1}^N \gamma_n (\alpha^n R + \beta^n S + C) \\ &= \sum_{n=1}^N (\gamma_n \times \alpha^n R + \gamma_n \times \beta^n S + \gamma_n \times C) \end{aligned} \quad (10.17)$$

Using Equation 10.17, the parameters  $\alpha$  and  $\beta$  in our fuzzy model in Equation 10.12 were obtained using the least squares method (Takagi and Sugeno, 1985) implemented

Table 10.4: Detailed tone pattern for sentences shown in Figure 10.4

Sentence ID	Sentence	Tone Pattern
S1	Ópé kíó tó dé sí ilé, nítorí ònà tó jìn ló ti rìn wá.	HHHHHHMH, HMHLLHMLH
S2	Ómòpé èmi kò ló gbé àga iyá àgbà, sórí igi emi.	HLHLMHHLMLHLL, HMMMM
S3	Bàbá àgbè tita kòkó rẹ, kótó mọpé kòkó ti gbówó lóri.	LHLLMMLHL, HHLHLMHHHH
S4	Òdòmi ló kòkó sáwá, kí ó tó rìn padà sí gbòngàn Ìbàdàn.	LLMHSHH, HHMLHLLLLL

in MATLAB 6.5. These parameters were obtained using the data described in Section 10.1.

In order to illustrate the behaviours of the input-output relation of the model with respect to the recorded data, we plotted the relation between the degree of lowering suffered by  $f_0$  contours of sentences for different tone combinations. Figure 10.4 provides a description of the relation between the degree of lowering and the position of the syllable in the sentences. It is noted that our model correctly captures the intonation phenomena discussed in *Connell and Ladd* (1990) and *Láníran and Clements* (2003). Each plot in Figure 10.4 depicts a twenty syllable sentence with two phrases. The detailed tone pattern for each sentence is depicted in Table 10.4.

The degree of lowering suffered by the last two syllables in Plot S4 increases drastically, with the  $f_0$  of the last syllable suffering more lowering. This demonstrates the well known final lowering intonation phenomenon, predominant in sentences ending with a sequence of L tone syllables, reported by *Connell and Ladd* (1990) and *Láníran and Clements* (2003). A similar general pattern is observed in Plot S3, but the degree of lowering towards the end of the sentence is much smaller with the last syllable increasing slightly. Sentences ending with mixed tone, plot S2, show similar patterns with those ending with a sequence of H tones but with less lowering and more stability at the end. Although the approximation provided by the relations in Figure 10.4 is rather coarse, the fuzzification together with global optimisation facilitated by the fuzzy model provides a continuous representation with the flexibility that is necessary to reproduce intonation contour with finer details.

We observed that the *BaseT* values are speaker dependent and we used a fixed

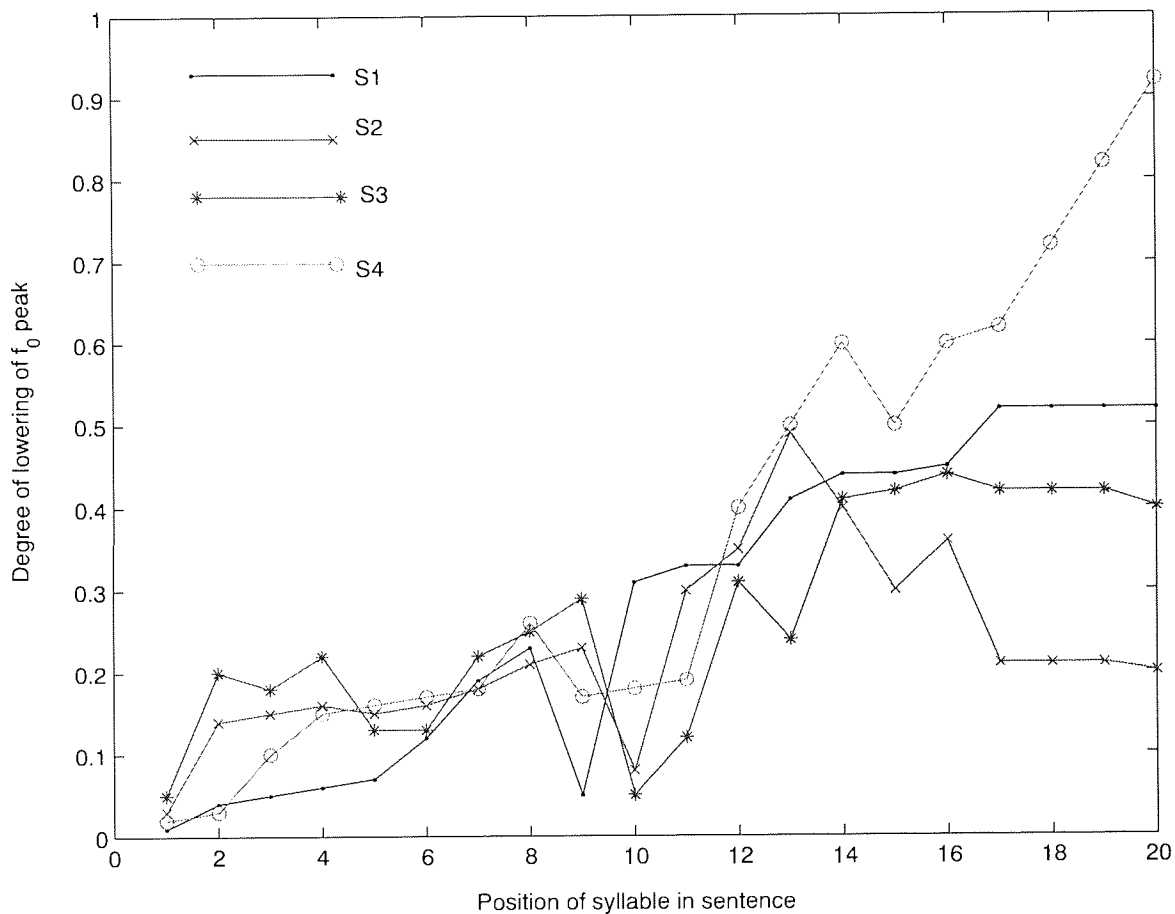


Figure 10.4: Relative Input-output behaviour as predicted by the model for different tone combination

value for each tone type. The value for the speaker used for developing the model is shown in Table 10.5. The variables *TonCon* and *RelPos* were computed from the data using the formula in Table 10.1. A set of input-output data of the form  $TonCon_j, RelPos_j \rightarrow f_{j+1}$  is then compiled, with  $j = 1, 2, \dots, 300$ . We have used these 300 cases to obtain the parameters for the fuzzy model in computing the  $f_0$  contour peak and valley respectively. In our earlier model (Ođéjóbí *et al.*, 2004b), fuzzy rules were used only to model the peaks of the  $f_0$  curve. The valleys were computed as a proportion of the realised  $f_0$  value of the peak. As a result, the model fails to capture important intonation phenomena such as final lowering. Hence, we now use fuzzy rules to model both the peak and the valley.

Table 10.5: Minimum  $f_0$  value perceptible as tone  $T \in H, M, L$  a speaker

Tone	Valley	Peak
	$f_0 Hz$	$f_0 Hz$
<i>BaseH</i>	85.00	88.00
<i>BaseM</i>	71.00	72.00
<i>BaseL</i>	60.00	63.00

The parameter identification process results in a total of ten rules each for computing the peak and valley of each tone type. Since we have a total of three tone types, there is a total of  $3 \times 2 \times 10 = 60$  rules in the fuzzy rule base. The parameters for each tone in these models are then obtained (Ođéjóbí *et al.*, 2004a). The resulting model captures the linguistic and speaker dependent features of intonation phenomena such as downstepping, final lowering,  $f_0$  reset, etc. It also generates an artificial intonation contour for any given target utterance. The input data needed for generating an intonation contour, i.e. syllable position and identity, are easily readable from text. This makes the incorporation of our intonation model to a practical TTS system a straight-forward task.

### 10.2.3 Generating intonation contour for an utterance

The  $f_0$  values for the peak and valley of a syllable in the intonation contour is computed using the algorithm shown in Figure 10.5.

We illustrate how the model computes intonation with the sentence depicted in Figure 10.7. The tone sequence in the 8-syllable sentence is: *H L H LM LLL*. The

```

Begin
   $k :=$  Total number of syllables in the utterance;
   $Sp :=$  An array for storing all syllables in a sentence;
   $N :=$  Total number of fuzzy rules in the model;
  For each  $p = 1, 2, \dots, k$ :
  {
     $Sp(p) :=$  ReadNextSyllable();
    If (Tone( $Sp(p)$ ) is not the first tone of type T) Then
    {
      Compute TonCon and RelPos for  $S(p)$ , using the
        formula in Table 10.1;
      Fuzzify the input data TonCon and RelPos, using
        the membership functions in Equations 11.4, 10.10 10.11;
      For each  $n = 1, 2, \dots, N$ :
      {
        Compute the premise and consequence values for the fuzzy rule  $R_n$ ;
        Compute the inference output of  $R_n$  using Equation 10.14;
      }
    }
  }
  Cubic Interpolation of points;
End

```

Figure 10.5: An algorithm for generating intonation contour for an utterance

canonical peak/valley of the syllables associated with these tone is shown in Table 10.6. Since the first two syllables, i.e. “Wón” and “sì”, are the first of their tones types, i.e. H and L respectively, they retain their canonical values (cf. Figure 10.5). The fuzzy rule is applied to the next syllable, i.e. “tún”, which carries an H tone with a canonical  $f_0$  peak and valley of 139.05 and 108.75, respectively. Using the formula in Table 10.1, the variable  $TonCon$  for the peak and valley are computed as  $101.10 - 139.05 = -37.95$  and  $88.31 - 108.75 = -20.44$ , respectively.  $RelPos$  for this syllable is  $\frac{3}{8} = 0.375$ . From Table 10.5,  $BaseT = BaseH = 88.00$ , we compute the following values for the peak of the syllable:

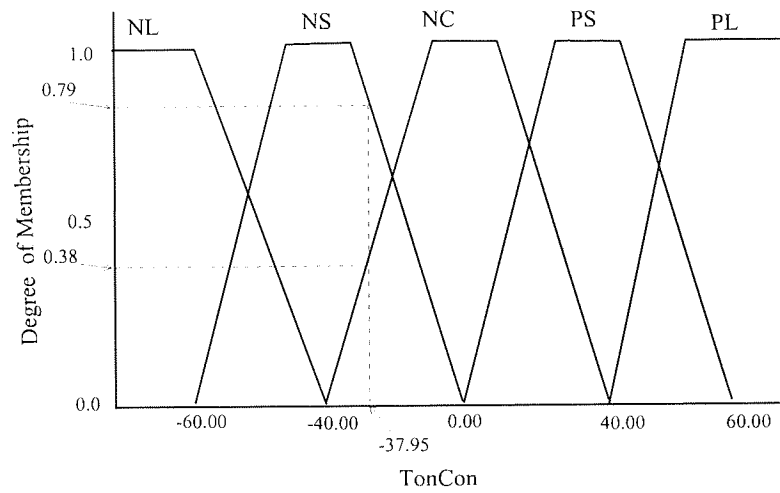
$$\begin{aligned}
 R &= \left( \frac{f_j^c}{f_{Last}^T} \right) \times (f_{Last}^T - BaseL) = \left( \frac{139.05}{150.01} \right) \times (150.01 - 88.00) = 57.48 \\
 S &= \left( \frac{f_{Last}^T}{f_i^R} \right) = \left( \frac{150.01}{101.01} \right) = 1.485 \\
 C &= BaseT = BaseH = 88
 \end{aligned}$$

In the fuzzification process, the degree of truth for each atom of the premise variable is determined using the membership functions for the respective variable. In the case of the  $TonCon$  value for the peak, we compute its membership value to each of the five fuzzy sets in  $A_j; j = 1, \dots, 5$ , i.e.  $\mu_{NL}(-37.95) = 0.0$ ,  $\mu_{NS}(-37.95) = 0.79$ ,  $\mu_{NC}(-37.95) = 0.38$ ,  $\mu_{PS}(-37.95) = 0.00$ , and  $\mu_{PL}(-37.95) = 0.00$ . For the  $RelPos$  value, the membership value is computed as  $\mu_N(0.375) = 0.58$  and  $\mu_F(0.375) = 0.47$ . Figure 10.6 depicts the graphical illustration of the computed membership functions<sup>1</sup>.

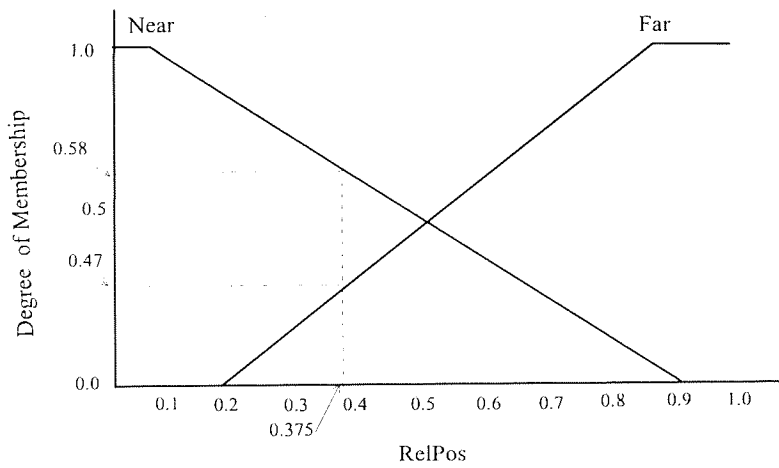
Column 2 in Table 10.7 shows how the computed degrees of membership are combined in the premise of the fuzzy rule using the minimum (min) operator (*Takagi and Sugeno, 1985*). Column 3 shows how the  $f_0$  value corresponding to the peak of the target syllable, i.e.  $\bar{f}_{i+1}$ , is computed in the consequence of each fuzzy rule. The premises and consequences of each rule are then used to infer the value of the peak of the syllable as follows:

$$\bar{f}_3 = \frac{\sum_{n=1}^{10} w_n \times f_{i+1}^n}{\sum_{n=1}^{10} w_n} = \frac{163.95}{1.41} = 116.28 \quad (10.18)$$

<sup>1</sup>Note that the figure has not been drawn to scale.



(a) Membership value for  $TonCon = -37.95$



(b) Membership value for  $RelPos = 0.375$

Figure 10.6: Computation of membership values for crisp input data

Table 10.6: Sentence data

Syllable ID	Position	Tone Type	Canonical	
			$f_0$ Peak (Hz)	$f_0$ Valley (Hz)
Wón	1	H	150.01	115.0
sì	2	L	101.09	88.30
tún	3	H	139.05	108.75
pò	4	L	85.06	59.08
we	5	M	95.11	70.34
Ì	6	L	78.55	60.41
bà	7	L	83.08	61.26
dàn	8	L	77.62	62.07

The computed points for all peaks and valleys for this sentence is shown in Figure 10.7. Figure 10.7(a) shows the S-Tree after the incorporation of the computed points. The corresponding waveform for this S-Tree is shown in Figure 10.7(b).

The abstract waveform generated by the S-Tree provides information on the perceptually significant points on the waveform which is necessary for the actual  $f_0$  computation. While the magnitudes of most of the points in the S-Tree shown in Figure 10.7 agree with the actual computed  $f_0$  values, this figure shows that the actual  $f_0$  values for peaks  $P_4$  and  $P_5$  do not correspond to the symbolically represented magnitudes on the S-Tree. This shows that the abstract waveform pattern generated by the S-Tree using phonological rules is not necessarily the same as what is observed at the phonetic level. This behaviour is supported by the fact that the synthesised  $f_0$  contour agrees with the natural  $f_0$  contour shown in Figure 10.8(b).

### 10.3 Model evaluation

We have carried out two types of evaluation on our model: (i) quantitative and (ii) qualitative. The quantitative evaluation examines how well the model predicts the data. This was achieved by calculating the Mean Square Error (MSE), Root Mean Square Error (RMSE) and Correlation of natural versus predicted outputs (*Hermes*, 1998; *Clark and Dusterhoff*, 1999). The qualitative evaluation examines the impression of native Yorùbá speakers with respect to the quality of the output in terms of intelligibility and naturalness. The intelligibility of our synthesised speech is evaluated using transcription error rate (*Wu and Chen*, 2001). For the evaluation of naturalness,



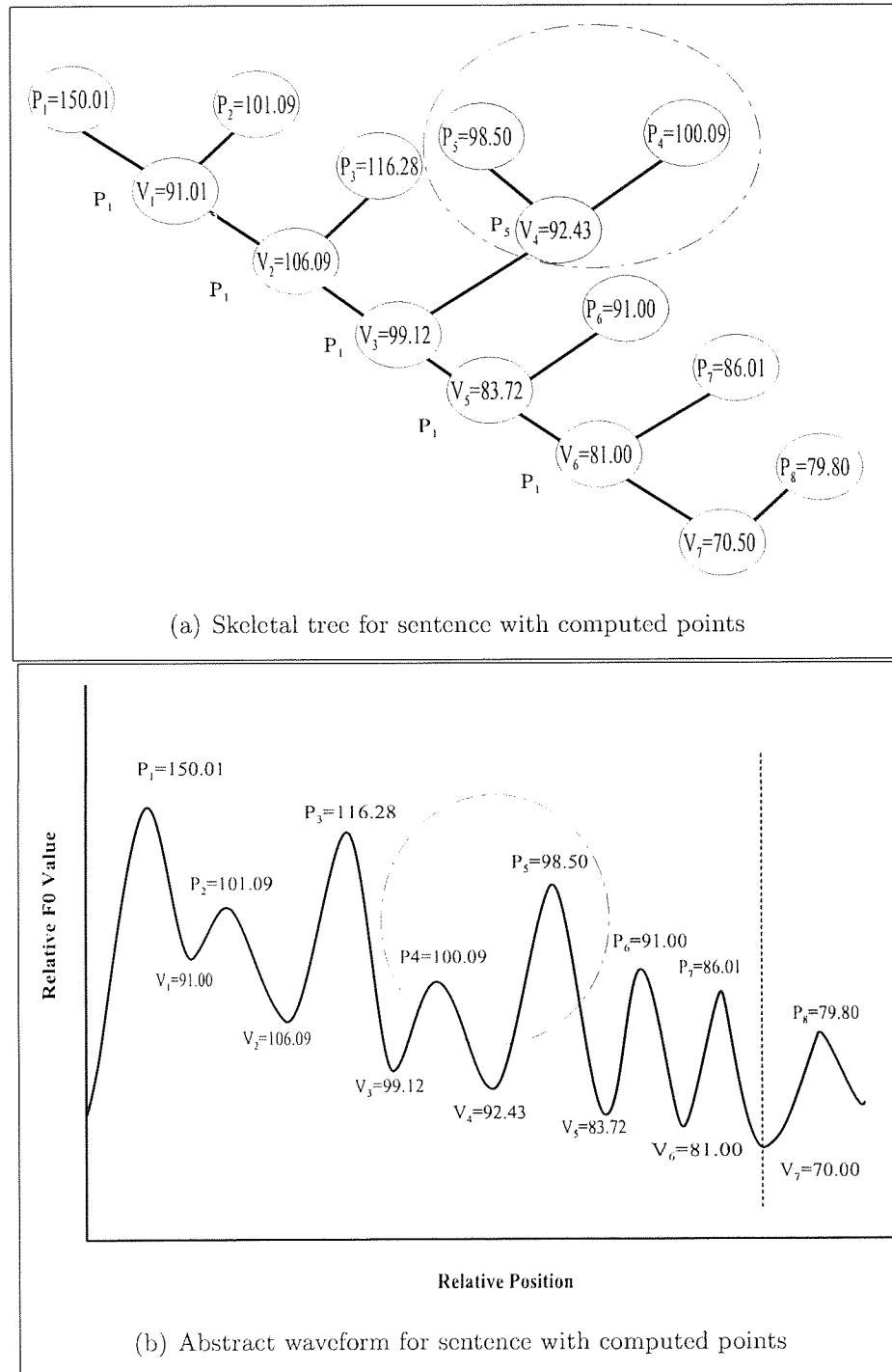


Figure 10.7: Computed points on the transcript of S-Tree generation for sentence “Wòn sù tún pòwe Ìbàdàn.”

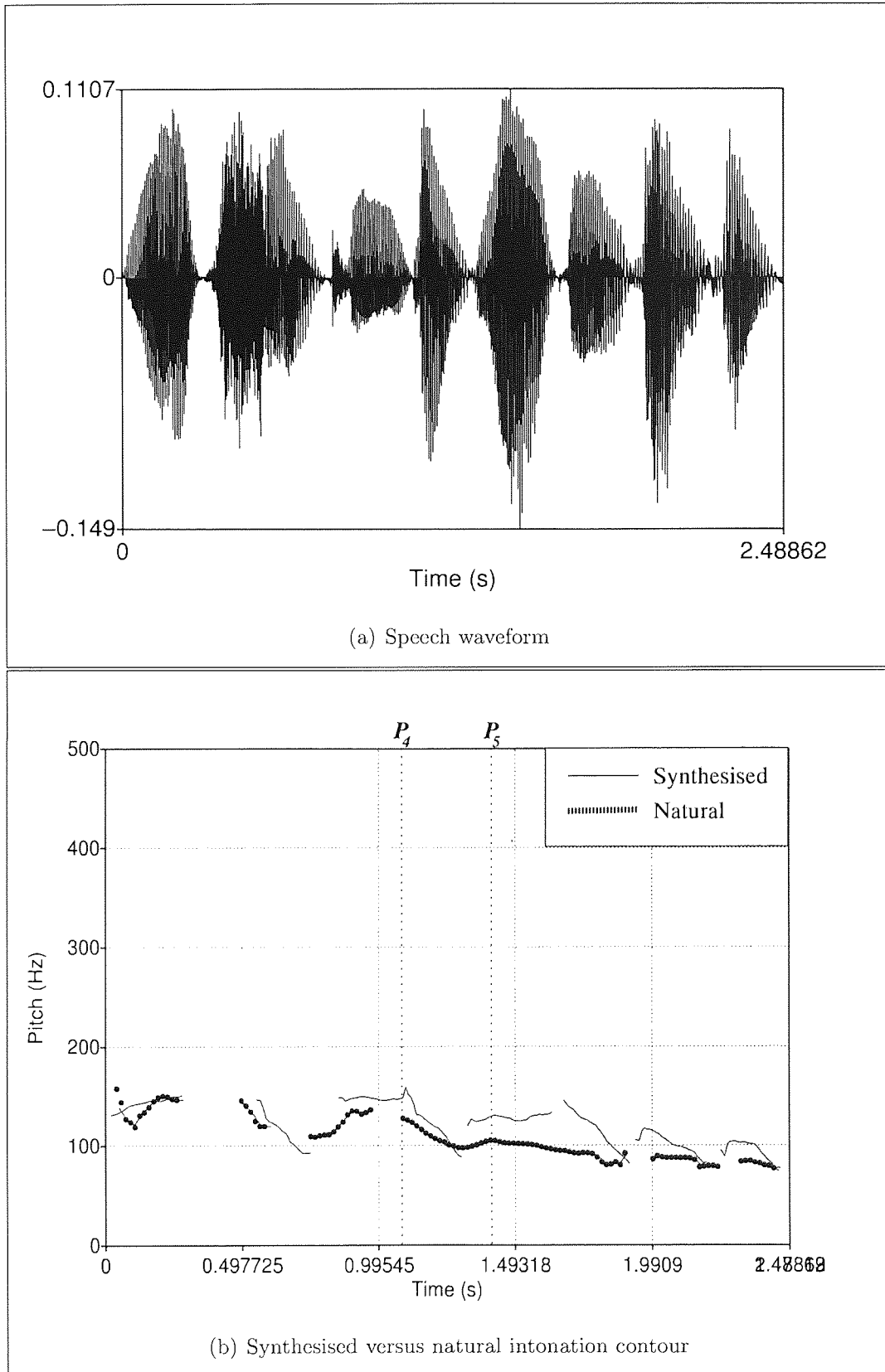


Figure 10.8: Natural versus Synthetic intonation of “*Won si tun powe Ibadan*”

Table 10.7: Premise and consequence computation for each rule

Rule No.	Premise	Consequence (Hz)
1	$\mu_{NL}(-37.95) \wedge \mu_N(0.375)$ $\min(0.0, 0.58) = 0.0$	$f_{i+1}^1 = 0.81R + 1.10S + 88.00 =$ 136.52
2	$\mu_{NS}(-37.95) \wedge \mu_N(0.375)$ $\min(0.79, 0.58) = 0.58$	$f_{i+1}^2 = 0.78R + 1.28S + 88.00 =$ 134.74
3	$\mu_{NC}(-37.95) \wedge \mu_N(0.375)$ $\min(0.36, 0.58) = 0.0$	$f_{i+1}^3 = 0.53R + 1.46S + 88.00 =$ 120.69
4	$\mu_{PS}(-37.95) \wedge \mu_N(0.375)$ $\min(0.0, 0.58) = 0.0$	$f_{i+1}^4 = 0.41R + 1.55S + 88.00 =$ 113.91
5	$\mu_{PL}(-37.95) \wedge \mu_N(0.375)$ $\min(0.0, 0.58) = 0.0$	$f_{i+1}^5 = 0.35R + 1.64S + 88.00 =$ 110.59
6	$\mu_{NL}(-37.95) \wedge \mu_F(0.375)$ $\min(0.0, 0.47) = 0.00$	$f_{i+1}^6 = 0.29R + 1.82S + 88.00 =$ 107.40
7	$\mu_{NS}(-37.95) \wedge \mu_F(0.375)$ $\min(0.79, 0.47) = 0.47$	$f_{i+1}^7 = 0.23R + 2.00S + 88.00 =$ 104.21
8	$\mu_{NC}(-37.95) \wedge \mu_F(0.375)$ $\min(0.36, 0.47) = 0.36$	$f_{i+1}^8 = 0.19R + 2.18S + 88.00 =$ 102.17
9	$\mu_{PS}(-37.95) \wedge \mu_F(0.375)$ $\min(0.0, 0.47) = 0.0$	$f_{i+1}^9 = -0.13R + 2.36S + 88.00 =$ 84.02
10	$\mu_{PL}(-37.95) \wedge \mu_F(0.375)$ $\min(0.0, 0.47) = 0.0$	$f_{i+1}^{10} = -0.21R + 2.55S + 88.00 =$ 79.69

we used the Mean Opinion Score (MOS) (*Donovan, 2003; Sakurai et al., 2003; Viswanathan and Viswanathan, 2005*) technique, with rating on a five point scale. The procedure and results of this evaluation is discussed in the following subsections.

### 10.3.1 Quantitative evaluation

Our quantitative evaluation was conducted using thirty SY statement sentences. Fifteen of the sentences contain syllables which form part of our training data set (cf. Section 10.1). The remaining fifteen sentences contain syllables that are outside our training data set. We have obtained separate MSE and RMSE values for our model using both the training set and the test set (cf. Tables 10.8 & 10.9).

To put our results in the context of related work, we have compared them with the results from other intonation models. Note that as the present work is the first on SY TTS prosody modelling, we can only compare our results with similar work available in the literature. Table 10.8 shows that the data-driven methods reported by

Table 10.8: Comparison of MSE results

Intonation Model	Language	MSE
<i>Elman-10</i> ( <i>Sakurai et al.</i> , 2003)	Japanese	0.074
<i>Tree-set-30</i> ( <i>Sakurai et al.</i> , 2003)	Japanese	0.068
Rule based ( <i>Sakurai et al.</i> , 2003)	Japanese	0.113
Our Model (with training set)	Yorùbá	0.095
Our Model (with test set)	Yorùbá	0.105

Table 10.9: Comparison of quantitative results

Intonation Model	Language	RMSE (Hz)	Correlation
Complete <i>RFC</i> ( <i>Taylor</i> , 2000b)	British English	14.60	0.651
Complete <i>Tilt</i> ( <i>Taylor</i> , 2000b)	British English	14.50	0.647
Our Model (with training set)	Yorùbá	15.56	0.612
Our Model (with test set)	Yorùbá	17.15	0.511

*Sakurai et al.* (2003) outperform our model in the quantitative analysis. For example, while our training and test sets produced an MSE of 0.095 and 0.105 respectively, the *Elma-10* and *Tree-set-30* neural network based models presented in *Sakurai et al.* (2003) produced a lower MSE of 0.074 and 0.068 respectively. However, our model outperforms the rule-based system presented in *Sakurai et al.* (2003), which has an MSE of 0.113.

In a further quantitative evaluation of our model, we compared the RMSE and correlation of our model with those from phonetically based intonation models, i.e. the Rise Fall Connection (RFC) and the *Tilt* model proposed by *Taylor* (2000b). Again, as shown in Table 10.9, the two methods produced better results than our model. These results suggest that the data-driven and phonetically based intonation models produced a better modelling of intonation data than our fuzzy logic based model. However, *Wang et al.* (2002), *Monaghan* (1990) and *Sakurai et al.* (2003) have shown that quantitative evaluation criterion does not necessarily express the perceptual quality of synthesised intonation contour. We therefore proceed to conduct a qualitative evaluation of our model.

### 10.3.2 Qualitative evaluation

Our preliminary qualitative evaluation involves a measure of how the perceptual quality of the synthesised speech mimics those of the natural speech in terms of the intelligibility and naturalness. The same training and test sets used for the quantitative evaluation were also used for the qualitative evaluation. To generate the synthesised utterances, we replaced the  $f_0$  curve of each syllable in the natural utterance with the ones generated by our intonation model using the *Praat* speech processing tool (Boersma and Weenink, 2004). The Pitch Synchronous Overlap (PSOLA) method was then used to synthesise the utterance (Moulines and Charpentier, 1990).

Nineteen adult native SY speakers were invited to participate in the qualitative tests. To ascertain their hearing ability, they were all subjected to an initial screening process. This process involves playing some natural speech sound to them and asking them to write down what they heard. Those who failed to produce 100% accuracy in this test were excluded from the evaluation experiment. Other participants were removed because their response were inconsistent. For example, some of them rated the quality of some synthetic speech higher than the natural speech. As a result, a total of seven participants were removed and twelve participated in the final qualitative evaluations.

### 10.3.3 Statistics difference calculation

We used the *sign test* (Anderson et al., 2002) to assess the statistical difference in the perceived quality of the synthetic speech produced by any two different models. The sign test is a non-parametric distribution free method of analysis for matched-pairs of data. An important feature of this test is that it does not require normality assumptions about the data. The sign test is very robust although insensitive (Rana et al., 2005). The computation of the sign test for pairs of sample data involves counting the number of positive differences between the matched pairs and relating them to a binomial distribution.

The aim of this test is to determine whether the synthetic speech generated using a method A is preferred by more listeners than the synthetic speech generated using another method B. Our sign test involves using a sample of  $m$  listeners to identify a

preference for one of two synthetic speech sounds generated using two different methods. Since our hypothesis is that method A produces a better synthetic speech than method B, we use a one-tailed test.

To do this, we first collect the independent sample scores for each sentence given by each listener. The average score for each sentence for all listeners is then computed for each of the model. To test the statistical difference of any two models, we pair their independent sample scores.

In the case of naturalness evaluation, we perform the sign test on our data by first collecting the independent sample MOS score for each sentence given by each listener. The average MOS for each sentence for all listeners is then computed for each of the models. To test the statistical difference of any two models, we pair their independent sample MOS scores for corresponding sentences. We omit pairs for which there is no difference so that we can have a possibly reduced sample of pairs. For example, if the MOS obtained for 30 sentences using model A is  $X = x_1, x_2, \dots, x_n$  and that for model B is  $Y = y_1, y_2, \dots, y_n$ , where  $n = 30$ , we collect the independent pairs of sample data for the population as  $(x_j, y_j); j = 1, \dots, 30$ . We then record the preference data for the average MOS of each sentence by using a plus sign (+) if the average MOS  $x_j$  is greater than  $y_j$  and minus (-) if the average MOS  $x_j$  is less than  $y_j$ . We ignore the situation when they are equal, i.e.  $x_j = y_j$ . The recorded preference data is used for our analysis. If  $w$  is the number of pairs for which  $y_i - x_i > 0$  and the null hypothesis,  $H_0$ , is true then  $w$  will follow a binomial distribution.

A similar process as described above is used in the case of intelligibility evaluation, except that in this case, the average intelligibility score for each listener, for each sentence computed by Equation 10.19 is used as the sample data for the methods to be compared.

Using these data, our objective is to determine whether there is a difference in preference between two speech sounds generated by the two methods being compared. If the null hypothesis,  $H_0$ , cannot be rejected, we will have no evidence indicating a difference in preference for the two methods. However, if  $H_0$  can be rejected, we can conclude that the listeners' preferences are different for the two methods. In that case, the method selected by the greater number of listeners can be considered the preferred

one.

### Intelligibility test

During the intelligibility test, only the speech sound with synthesised  $f_0$  contour was played to each participant. After a speech sound is played, the participant is asked to write down what they heard. Our intelligibility evaluation is very rigorous, in that the participants were not only required to identify the tones on the synthesised utterance, they were also required to accurately identify the syllables associated with each tone. Equation 10.19 is used to compute the transcription error rate.

$$\text{Intelligibility} = \frac{T_{All} - T_{Wrong}}{T_{All}} \times 5.0 \quad (10.19)$$

where  $T_{All}$  is the total number of syllables in a sentence and  $T_{Wrong}$  is the number of syllables that had been wrongly transcribed. The value 0 will be obtained from the computation if a participant transcribes all the syllables in the synthesised utterance wrongly. The value 5 is obtained when all the syllables in the utterance are correctly transcribed by the participant. Using this scale, the average value of 4.75, with a standard deviation of 0.56, was obtained for the training set. For the test set, we obtained a value of 4.03, with a standard deviation of 0.62. A sign test statistics shows that the training set is significantly ( $p \leq 0.05$ ) more intelligible than the test set sentences. The intelligibility rating of our model (from both training and test sets) compares well with that of other models listed in Table 10.10.

Table 10.10: Comparison of intelligibility results

Intonation Model	Language	Intelligibility <sup>†</sup>	Percentage
<i>Lee et al.</i> (1989)	Mandarin	96.00 (3.60)/100.00	96%
<i>Wu and Chen</i> (2001)	Mandarin	96.90/100.00	96.90%
Our model (with training set)	Yorùbá	4.75 (0.56)/5.00	94.00%
Our model (with test set)	Yorùbá	4.03 (0.62)/5.00	80.60%

<sup>†</sup>Numbers in parenthesis indicate standard deviation.

Numbers after '/' indicate scale of evaluation.

Table 10.11: Qualitative evaluation scores for naturalness

Value	Description
5	Perfect, indistinguishable from natural speech quality
4	Very good
3	Average
2	Poor
1	Weak or not acceptable

### Naturalness test

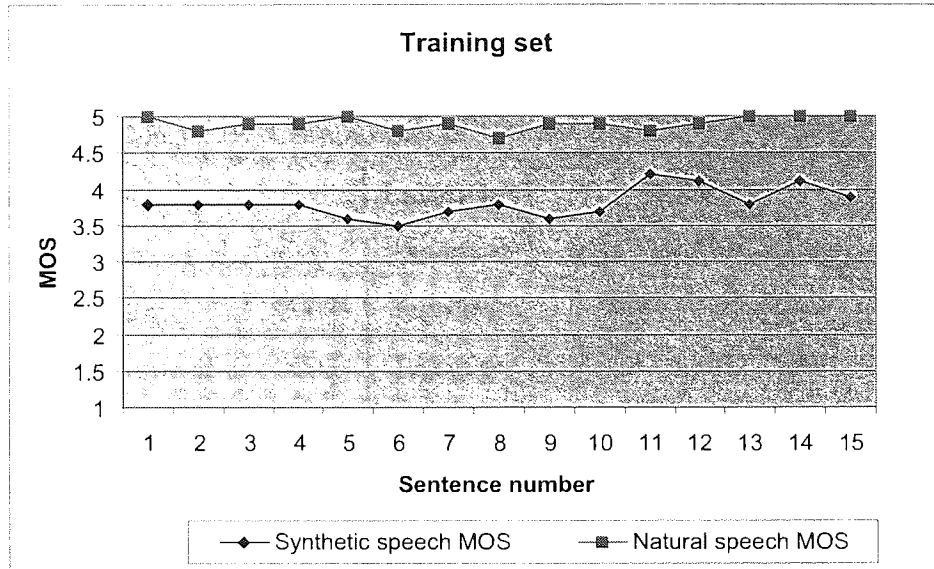
We used the Mean Opinion Score (MOS) (Monaghan, 1990; d'Alessandro and Mertens, 1995; Sakurai et al., 2003) to evaluate the naturalness of our synthesised speech. In carrying out the naturalness evaluation, we used two kinds of stimuli: modified and unmodified (Sakurai et al., 2003). The unmodified stimuli are naturally produced utterances recorded without any modification to the acoustic data. The modified stimuli are versions of the same naturally produced utterances whose the  $f_0$  data have been replaced by those generated by our model. The  $f_0$  manipulation for the modified stimuli was achieved using the PSOLA technique in *Praat*.

During the naturalness test, the modified and unmodified samples were randomly presented to each of our participants. They were not informed which sample was modified or unmodified. After each sample was presented, the participants were asked to rank their overall impression of the presented utterance using the five point scale shown in Table 10.11. The average MOS of each sentence for the training and test sets are plotted in Figure 10.9(a) and Figure 10.9(b) respectively.

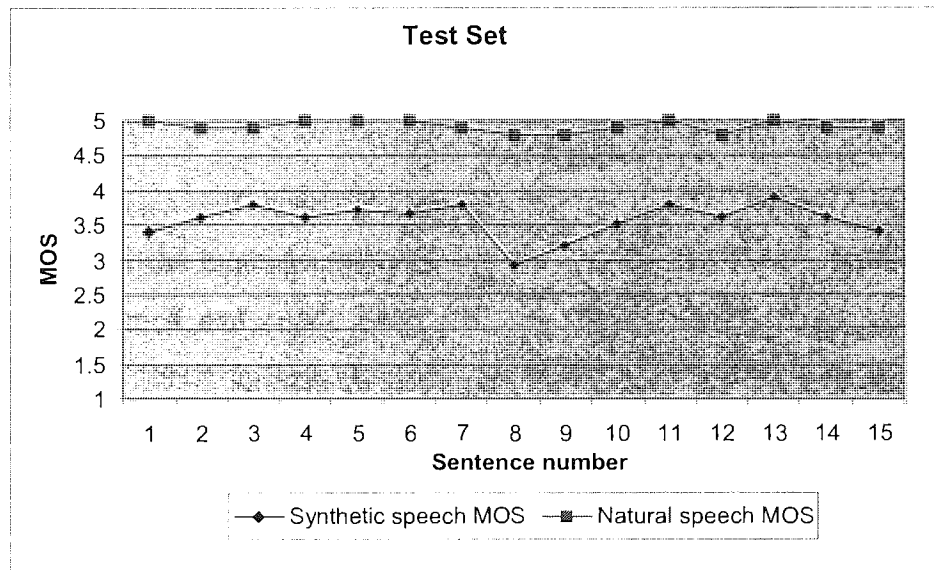
The results show that the participants consistently ranked the natural speech high with an overall average MOS of 4.910. A sign test statistics indicates that there is a strong evidence ( $p \leq 0.001$ ) that the natural speech is preferred over the synthetic speech generated using the training set. The overall average MOS for the training and test sets is 3.813 (0.196) and 3.565 (0.260) respectively. However, there is no statistically significant evidence ( $p > 0.05$ ) that the speech generated using the training set is preferred over that generated using test set.

When put in the context of the results obtained for similar intonation models in the





(a) MOS for the training set



(b) MOS for the Test set

Figure 10.9: MOS results for the naturalness test

literature, the MOS scores for the naturalness (cf. Table 10.12) show that our model has the potential to outperform the rule-based models presented in *Sakurai et al.* (2003) & *Mittrapiyanuruk et al.* (2000) and the data-driven Elman-10 model presented in *Sakurai et al.* (2003). The results from our model also compare well with other state-of-the-art models (e.g. *Lee et al.* (1989); *Wu and Chen* (2001)).

Table 10.12: Comparison of the naturalness test results

Intonation Model	Language	Naturalness <sup>†</sup>	%
Elman-10 ( <i>Sakurai et al.</i> , 2003)	Japanese	3.240 (0.722)/5.000	64.80%
Tree-set-30 ( <i>Sakurai et al.</i> , 2003)	Japanese	3.690 (0.746)/5.000	73.80%
Rule-based ( <i>Sakurai et al.</i> , 2003)	Japanese	2.830 (0.934)/5.000	56.60%
<i>Lee et al.</i> (1989)	Mandarin	89.300 (8.100)/100.000	89.30%
<i>Wu and Chen</i> (2001)	Mandarin	3.600/5.000	72.00%
<i>Mittrapiyanuruk et al.</i> (2000)	Thai	3.100/5.000	62.00 %
Our model (with training set)	Yorùbá	3.813 (0.196)/5.000	76.26%
Our model (with test set)	Yorùbá	3.565 (0.260)/5.000	71.30%

<sup>†</sup>Numbers in parenthesis indicate standard deviation. Numbers after ‘/’ indicate scale of evaluation.

We observed also that, generally, the transition of the  $f_0$  contour (cf. Figure 10.8) is consistent with the linguistic descriptions reported in respect of Yorùbá intonation phenomena (*Connell and Ladd*, 1990; *Láníran and Clements*, 2003). Although the perception of synthetic speech generated by a given TTS system depends on several factors including: (i) the properties of the speech signal modelled, (ii) language being synthesised, (iii) the experience of the participants involved in the qualitative evaluation, and (iv) the context of the test materials (*Aker and Lennig*, 1985), our preliminary results nonetheless show that the proposed intonation model is creditable when compared with other models.

Recall that our model uses linguistic terms to describe intonation phenomena which is close to how they are being described by linguistic experts. This makes our model easier to comprehend when compared with data-driven or other rule-based models. Another strength of our model is that it can be used to explore the relationship between

speech data and linguistic knowledge. This relationship will help to determine the aspects of speech that are perceptually more significant. Such information can then be used to direct the data collection process in data-driven approaches.

## 10.4 Summary

This chapter presents the intonation modelling approach and demonstrates its applicability using the Standard Yorùbá language. We performed both quantitative (Mean Square Error, Root Mean Square Error and Correlation) and qualitative (Mean Opinion Score) evaluations of our model. The results suggest that, although the model does not predict the numerical speech data as accurately as contemporary data-driven approaches, it produces synthetic speech with a better naturalness rating.

In the next chapter, we shall present the duration model. The duration model which predicts the most important dimension of prosody in our overall prosody model, i.e. the duration dimension. Our proposed duration modelling approach uses the fuzzy decision tree technique.

# Chapter 11

## Duration modelling

Timing is an important cue in the perception of coherent speech. The duration dimension is therefore an important component of speech prosody. *Chung and Seneff* (1999) suggested that listeners make linguistic decisions on the basis of durational cues which can serve to distinguish, for example, phrase-final versus non final syllables. The function of duration modelling in a TTS system is to compute the time interval that will be occupied by each speech unit <sup>1</sup> in the overall duration of an utterance. In addition, the duration information is also used to align the other dimensions of speech, e.g. the  $f_0$  and intensity dimensions, to the segmental structure of the syllable. The accurate prediction of this time interval from text in a TTS synthesis system remains a difficult task for a number of reasons.

A review of the literature on the state of the art in speech technology (*Chung, 1997; van Santen et al., 1997; Vainio, 2001*) suggests that the duration patterns of speech and the many sources of variability which affect them is still not well understood. However, it is well known that duration values are associated intrinsically with individual phones but are altered by a number of contextual factors. These factors are also known to interact with each other in a complex way. Developing a database in which all the possible factors in all possible combinations are adequately represented is a difficult, if not an impossible, task. The high number of factors affecting duration makes their interactions obscure one another rendering their presence difficult to quantify and analyse.

---

<sup>1</sup>in this work, we consider a syllable as a speech unit

In order to account for most of the complex behaviour of the factors that affect duration, it is important to augment numerical data with linguistic information derived from the observation of duration phenomena in speech sound. For example, it is well known that the duration of syllables at the end of an utterance increases due to the well known final lengthening phenomena. Although statistical and numerical data have been used to describe this phenomenon, the relationship between such increases or decreases in duration and the phonetic features of the syllables are best described linguistically.

We have explained in Chapter 6 that the Fuzzy Decision Tree (FDT) is an appropriate tool for modelling and integrating numerical and linguistic descriptions. FDT has been applied to the modelling of various problems, such as power system security assessment (*Boyer and Wehenkel, 1999*), weather forecasting (*Yuan and Shaw, 1995; Dong and Kothari, 2001*) as well as to software quality models (*Pedrycz and Sosnowski, 2001*). *Suárez and Lutsko (1999)* have assessed the performance of the FDT technique in real world problems such as classification of diabetes data, breast cancer data, heart diseases data as well as in waveform recognition.

In *Mitra et al. (2002)*, FDT was applied to the recognition of vowels produced by a group of male speakers in a Consonant-Vowel-Consonant context. The results of these applications suggest that the FDT technique is better in extrapolating from training data when compared with binary decision trees such as the Classification and Regression Tree (CART). A survey of the literature on duration modelling (e.g. *Chung and Huckvale (2001); Batůšek (2002)*), in the context of prosody modelling for TTS applications, suggests that CART is the most frequently used modelling technique. There is no reported work on the application of FDT to duration modelling.

In this Chapter we will first illustrate the application of the Fuzzy Decision Tree (FDT) technique to duration modelling in the context of SY prosody modelling. We will then compare the results of the FDT-based model with that of a CART-based duration model developed using the same speech database. We will also demonstrate these duration models within the context of our R-Tree based prosody modelling.

## 11.1 A preliminary analysis of factors affecting duration in SY

To establish a suitable focus for our experiments, we conducted some informal experiments on the speech database discussed in Chapter 8. We found that the acoustic duration of syllables in sentences varies considerably from their canonical duration. We have observed that female speakers show more variation in speech duration pattern and that it is difficult to locate the peak or valley in the  $f_0$  pattern of the female citation syllables. Such high variations in female speech data have also been reported in *Klatt and Klatt* (1990). The variations in the duration data of female voices in our speech database were too high to allow an objective analysis and modelling. The male speech data on the other hand has more stable duration and  $f_0$  pattern. We therefore base our analysis on the male speakers' data only.

The following linguistic factors were hypothesised to affect the duration of syllables in our informal experiment: (i) the phonetic structure of the target syllable, (ii) the phonetic structure of the neighbouring syllables (immediate left and right context only), (iii) the tone carried by the syllable, (iv) the position of the syllable in the word, (v) the length of the word, and (vi) the position of the word containing the syllable in sentence.

Among other features, our analysis clearly shows that syllables at the end of a sentence tend to be longer. We also observed that the length of the last syllable also increases by as much as 15%, if it is carrying a low tone; and 5% and 7% when carrying the High or Mid tone respectively. Based on our observations and the information derived from the literature as summarised above, we selected nine factors for analysing our duration data. If we represent the target syllable (i.e. the syllable for which duration is to be computed) as  $S_{tag}$  and the word in which it occurs as  $W_{tag}$ , the factors affecting duration that we selected for our duration data analysis are described as follows:

1. the position of syllable in  $W_{tag}$ ,  $S_{tag}^{PoW}$
2. the position of  $W_{tag}$  in the sentence,  $W_{tag}^{PoS}$
3. the length of  $W_{tag}$ ,  $W_{tag}^{len}$ , calculated as the number of syllables that it contains
4. the peak of the stylised  $f_0$  curve on  $S_{tag}$ ,  $S_{tag}^{f_0}$

5. the phonetic structure of the target syllable  $S_{tag}$ , i.e.  $S_{tag}^{pho}$
6. the phonetic structure of the preceding syllable,  $S_{pre}, S_{pre}^{pho}$
7. the phonetic structure of the following syllable  $S_{fol}, S_{fol}^{pho}$
8. the peak of the  $f_0$  curve on the preceding syllable  $S_{pre}, S_{pre}^{f_0}$
9. the peak of the  $f_0$  curve on the following syllable  $S_{fol}, S_{fol}^{f_0}$

### 11.1.1 Duration factor level and hierarchy

We employed a systematic approach to analyse the confounding factors on duration by first classifying the factors affecting duration into four levels:

- Level 0: syllable level,
- Level 1: syllable context level,
- Level 2: word level, and
- Level 3: word context level.

To arrive at this classification, we assume that, although the factors affecting duration are confounded, it is possible to analyse them hierarchically. This hierarchical view allows us to conceptually isolate factors operating at the syllable level from those operating at higher levels, e.g. at the word level. For example, the intrinsic properties of the syllable, such as the phonetic feature of the base syllable and the tone associated with the syllable can be conceptually separated from the contextual factors due to the position of the syllable within a word. In this respect, if we know the acoustic properties of the citation syllable and how they relate to its duration, we can compare this with the duration of the syllable in a word context. In this way it becomes possible to isolate parameters such as syllable phonetic properties and tone from higher level properties such as the position of the syllable in a word.

The Level 0 factors are primary factors inherent in the canonical syllable. They can be analysed using the acoustic properties of citation syllables. They can therefore be viewed as the atomic properties of the syllable which account for its primitive durational pattern. Factors affecting duration at Level 1 are confounded with Level 0 factors. For example, the contribution of the effects of a preceding syllable on a target syllable can

only be determined after we have isolated the duration of the target syllable based on its tone and syllabic structure. The Level 2 factors are confounded with both the Level 0 and Level 1 factors. For example, we cannot estimate the contributions of the effects of the position of a syllable in the word — which is a Level 2 factor — if we do not know the effect of the preceding, and possibly the following, syllable on its duration. The effect of the preceding and following syllables, which are level 1 factors, cannot be isolated from the Level 0 factor. The same explanation goes for the Level 3 factors. In Table 11.1, the duration factors in level  $n$  subsume those in level  $n - 1$ .

Our current analysis does not take inter-sentential factors into account. This is because our current focus is on duration modelling in the context of SY speech synthesis prosody for a sequence of isolated sentences. In this context, sentences are usually treated as the basic unit of structure (*Smith, 2004*). We view a paragraph as a sequence of sentences. The goal of our analysis is to use the results in training our duration model to learn the interaction between the identified duration affecting factors automatically and use them to predict the duration of syllables in the context of TTS.

Table 11.1: Levels of syllable duration affecting factors

LEVEL	DESCRIPTION	FACTORS
0	syllable	(1) phonetic structure, (2) the $f_0$ of the syllable
1	syllable context	(1) the $f_0$ of the preceding syllable, (2) the $f_0$ of the following syllable, (3) phonetic structure of the preceding syllable, (4) phonetic structure of the following syllable
2	word	(1) length of word, (2) position of syllable in word
3	word context	(1) position of word in sentence

## 11.2 Statistical analysis and observation

The analysis of the text database informed the selection of the text for our speech corpus. Out of the 690 SY syllables (cf. Table 3.3), we selected 456 syllables. These syllables are carefully selected to reflect the coverage of all syllable types in terms of phonetic and phonological distributions. For example, in the CV syllable type, the manner of articulation of the onset is considered. The onset consonants are selected



from each manner of articulation classes, i.e. stop, labio-velars, fricates, affricate, sonorants or semivowels. The selected onset is combined with each vowel type, e.g. Close rounded, Half-closed front, etc., in order to select the syllable for each class of utterance. The same process is repeated for all the syllable types. The data set adequately represents all SY syllable types (i.e.  $CV$ ,  $CVn$ ,  $Vn$ ,  $V$  &  $N$ ).

The 95 sentences in our corpus contains 350 different syllables. Sixty sentences were selected from our corpus for training our duration model. Forty of them are one-phrase sentences and each of the remaining twenty sentences contains two phrases. For the test data set, we have chosen thirty sentences from the remaining thirty-five sentences in our corpus. Twenty-one of them are one-phrase sentences and the remaining nine are two-phrase sentences. The distribution of syllable and tone types in these sentences is shown in Table 11.2.

Table 11.2: Statistics for the characteristics of the training and test data sets

Category	Description	Training set	Test set
Sentences	One-phrase	40 (60%)	21 (67%)
	Two-phrase	20 (40%)	9 (33%)
Words	Total word count	793	386
Syllable types	$CV$	1007	414
	$CVn$	368	234
	$Vn$	291	58
	$V$	755	292
	$N$	38	20
Tone types	$H$ tone	984	458
	$L$ tone	980	305
	$M$ tone	492	255

Both the training and test data sets contain semantically well-formed statement sentences and they are selected to reflect common, everyday use of SY. Within the training data set, the minimum number of syllables per sentence is 6 and a maximum of 24 syllables. The H and L tone syllables account for 40% each while the M tone syllables account for the remaining 20%. For the test data set, the minimum number

of syllables per sentence is 4 and a maximum of 19 syllables. The H, L and M tone syllables account for 45%, 30% and 25% each, respectively. Our training and test data sets contain statement sentences only because research on SY language intonation has shown that the mode of the sentences does not affect the intonation (*Connell and Ladd, 1990*). Since intonation has the closest proximity to sentence mode, we assume that its effects on the other dimensions of prosody, i.e. duration and intensity, if present, will be minimal.

To obtain the timing information of each syllable in context for our duration modelling and evaluation experiments, we have recorded and annotated all of these 90 sentences.

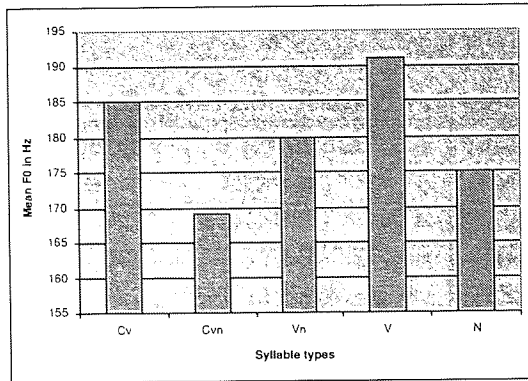
### 11.2.1 Level 0 factors

In our analysis, Level 0 is concerned with the observable factors affecting duration at the syllable level. Our analysis focused on how the phonetic structure and the tone of an SY syllable contributes to the estimate of the duration of that syllable.

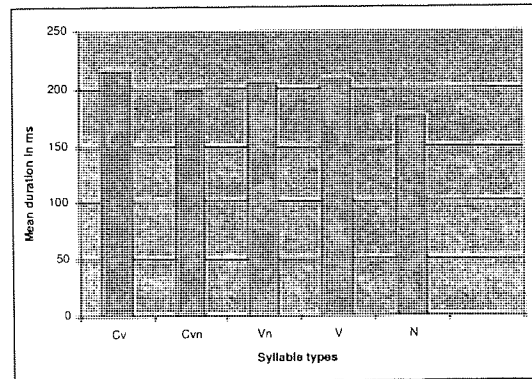
#### Phonetic structure of syllables

It has been suggested that syllable duration increases with the number of constituent phonemes in the syllable (*Chen et al., 2003*). If this suggestion applies to SY, a typical CV $n$  syllable would generally be longer than a CV syllable, and a CV syllable would be longer than a V or an N syllable, provided that they all carry the same tone. However, the  $n$  in the SY CV $n$  syllable is purely an orthographic device, hence the CV $n$  and the CV syllable types have the same number of segments. However, as shown in Figure 11.1, CV syllables are generally longer than their CV $n$  counterparts. For example, in Figure 11.1b, our CV syllables carrying the H tone have a mean duration of 215ms ( $SD$  35), which is longer than the corresponding CV $n$  syllables with a mean duration of 198ms ( $SD$  33). In addition, our data reveal that V syllables are longer than V $n$  syllables. It is important to note that V $n$  syllables are the nasalised equivalent of corresponding V syllables. In general, the syllabic nasals, i.e. N syllables, are the shortest syllables.

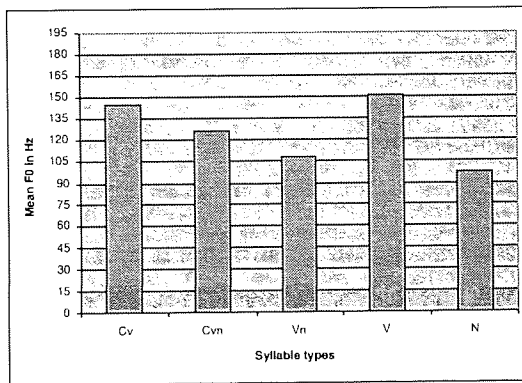
We also observed that SY syllables bearing a fricative consonant (i.e. an onset



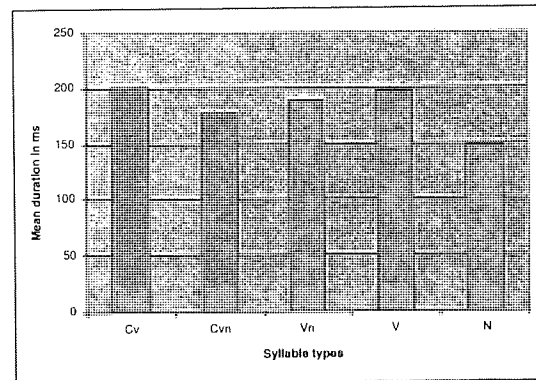
(a) Mean of H tone  $f_0$



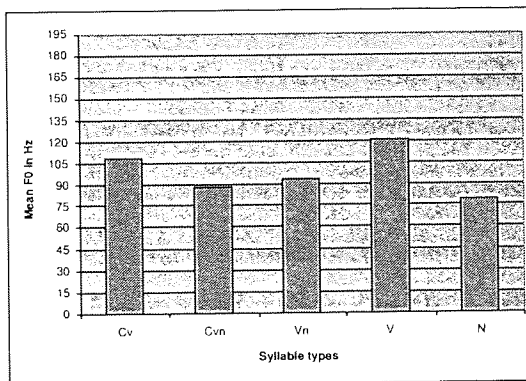
(b) Mean of H tone duration



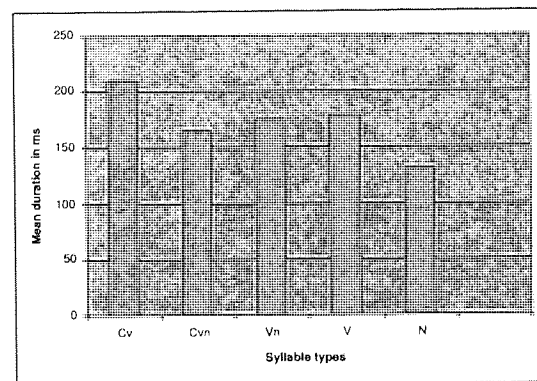
(c) Mean of M tone  $f_0$



(d) Mean of M tone duration



(e) Mean of L tone  $f_0$



(f) Mean of L tone duration

Figure 11.1: Level 0 Factors on syllable duration

represented by the grapheme *f*, *s* or *ʃ*, cf. Figure 3.1) tend to be longer than other syllables. This observation is consistent with that observed in Korean (*Chung and Huckvale, 2001; Huckvale, 2002*). Table 11.3 shows the statistics of the duration of various types of SY syllables.

Table 11.3: Statistics of the duration of various SY syllable types

Syllable type	minimum (msec)	maximum (msec)	Mean Duration (msec)	Standard Deviation (SD)
CV	180.59	235.89	216.15	42.83
V	172.24	230.34	197.26	22.34
V <sub>n</sub>	112.40	215.22	200.34	28.12
N	102.10	201.12	149.14	10.03
CV <sub>n</sub>	152.56	239.33	178.29	19.75

### Syllable tone

The tone of a syllable is characterised by the shape of the  $f_0$  curve, duration and intensity. However, the  $f_0$  curve provides a more potent way of describing the tonal signature of a syllable (*Connell, 2004*). We have shown in Chapter 9 that the turning points (i.e. peak and valley) on a stylised  $f_0$  curve of a tone carry perceptually significant information in respect of the tone. The duration of a syllable and the point in time when the peak and valley occur are also crucial components in the intelligibility and naturalness of a tonal signature (*Fagyal, 2002; Li, 2004*).

*Bákàrè (1995)* and *Harrison (2000)* are two important studies on duration of SY tones. The focus of *Bákàrè's (1995)* work is acoustic and perceptual and he recorded two speakers' production of 90 citation SY syllables. *Bákàrè* found that H tone syllables have the shortest duration. He also found that the duration of M tones are longer than that of L tones. *Harrison (2000)*, working with 20 citation SY syllables, found that the duration of mid- and low-toned syllables varies very slightly between 205ms and 220ms, while that of high-toned syllables is shorter at an average of about 160ms. *Harrison's (2000)* result on H tone duration is similar to that of *Bákàrè*. In respect of the M and L tones, however, *Harrison* found that there is no significant difference in the duration of M and L tones. Both studies employ similar approaches in the recording of the speech data for their experiments.

Our measure of duration differs from those reported above in two ways. First, our tone data is based on the numerical values of the peak of the  $f_0$  curves on syllables rather than the phonological tone as done in the other works. Second, our analysis is based on the stylised  $f_0$  of 257 SY syllables. We observed that the longest tone is the H tone (with a mean value of 237.70ms (SD 31.50)), while the shortest is the L tone (with a mean value 170.43ms (SD24.56)). The duration of the mid tone (221.38ms, (SD 29.34)) is longer than that of the L tone but shorter than that of the H tone. These results do not totally agree with the results of the two studies discussed above. The difference between our results and those discussed above may well be due to speaker-specific characteristics.

An analysis of the interpolation of 3<sup>rd</sup> degree polynomials into the  $f_0$  curves of the three SY tones (cf. Chapter 9) shows that the configuration of SY high tone is similar to that of Mandarin Tone-2 as reported by *Xu* (1997). The configuration of SY mid tone is similar to Mandarin Tone-1 and that of SY low tone is similar to Mandarin Tone-4. Since Tone-2 is longer than Tone-1 and Tone-1 is longer than Tone-4, we see that *Xu's* (1997) result is consistent with ours if we interpret the result in the context of  $f_0$  pattern. For example, we observed from our experimental data that syllables with high tones are, in general, longer than low tone syllables. However, the relative duration occupied by H-toned syllables is also contingent upon the identity of phonemes within the syllables.

### 11.2.2 Levels 1–3 factors

In our analysis, Levels 1–3 factors deal with contextual factors which cannot be attributed to the intrinsic properties of the syllable. These contextual factors range from syllable in word, phrase or sentence context. Our current analysis does not take inter-sentential factors into account.

#### Syllable context

The duration pattern of a number of syllables selected based on their types and tones were observed using carrier words configured as  $S_a S_t S_b$ ,  $S_t S_a$  and  $S_a S_t$ , where  $S_t$  is the target syllable and  $S_a$  and  $S_b$  are syllables with different tonal and phonetic types

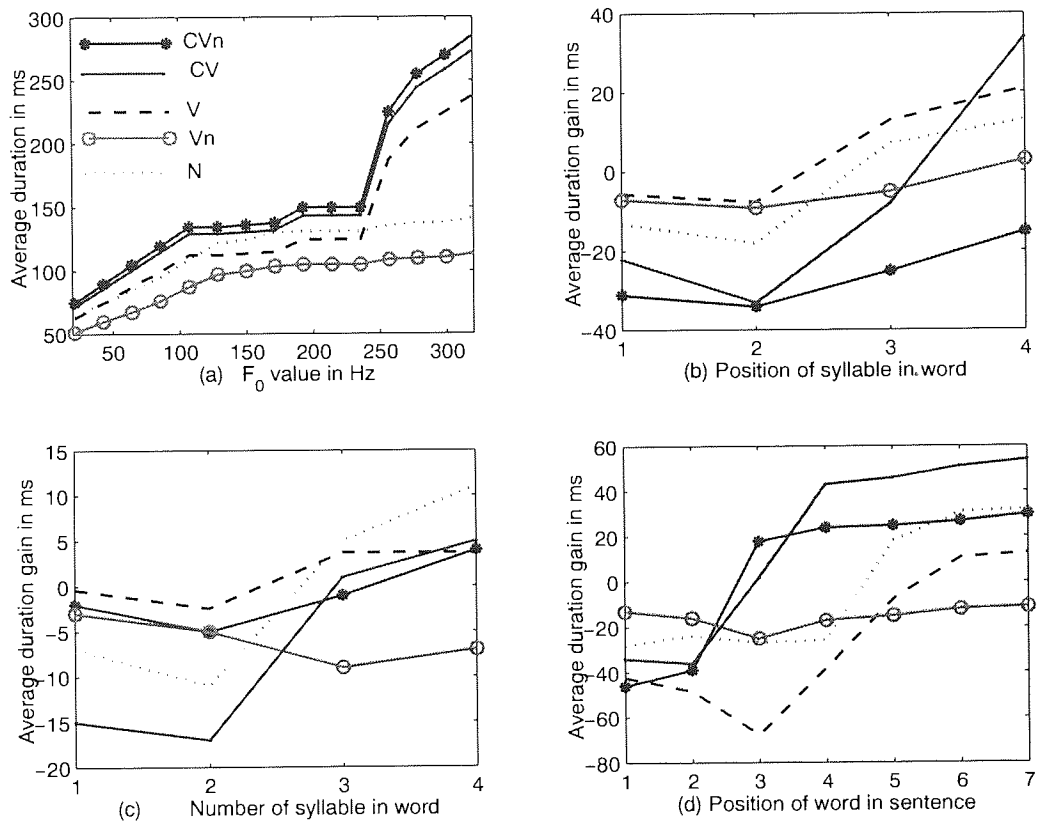


Figure 11.2: Factors affecting syllable duration in SY utterance as observed in the training corpus

preceding and succeeding  $S_t$  respectively. The aim of constructing these carrier words is to determine the contextual effects of the syllable on its duration. We did not observe any effect from the  $f_0$  values of the preceding and following syllables on the duration of the target syllable. We observed that there appears to be a tendency for syllables to be shortened when they are preceded by a much longer syllable. This observation is however insignificant and the range of values we obtained was too unstable for making a generalised conclusion.

### Word length

Figure 11.2c shows the relationship between duration and the length of a word. The length of a word is measured as the total number of syllables in the word. The general pattern is that as the number of syllables increases, the average syllable duration decreases. As shown in Figure 11.2c, this linear relation, with negative slope, is observable in words up to 3 syllables long. The relation seems to increase for words with 4 syllables and the slope of the relation changes to positive resulting in an increase in the average duration. On a close examination of our data, we found that a substantial part of this longer durations may have been as a result of syllables in the last positions in the word. The last syllable accounts for about 13% of the average increase in length.

### Position of syllable in word

On close examination of our data (as shown in Figure 11.2b), we found that syllables in the anti-penultimate, penultimate and last syllable position in a word have longer duration. We interpolate a 3<sup>rd</sup> degree polynomial into the duration data as a function of position (*Yang*, 1998). We obtained an  $R^2$  value of 0.13, for syllables in the medial position, and 0.11, 0.25 and 0.65 for syllables in anti-penultimate, penultimate and last positions respectively. This result indicates that the word length factor accounts for about 65% of the duration of syllables in the final position, while it accounts for about 11% of the syllable's duration in the initial position. We observed that L tone syllables in the final position undergo greater lengthening.

### Position of word in sentence

As shown in Figure 11.2d, all words in the initial and medial position are considerably shorter than the corresponding words in the final position. Medial words are slightly longer than initial words. We observed a lengthening of syllables with respect to the position of the word in which they are present in the sentence. This result suggests that the lengthening of words towards the end of a sentence results in the corresponding lengthening of their component syllables.

### 11.2.3 Pause duration

Based on the duration data of our annotated sentences, we obtained the duration of pauses and their locations within SY utterances. The pause duration within a word is zero due to the overlapping of syllables in spoken words. Between words, pause duration depends on the location of the word in a sentence. The duration between words generally increases towards the end of the sentence. Pause durations also result from punctuation marks appearing in a sentence. The pause duration for comma (,) has a mean value of  $97ms$  ( $SD$  12.3).

## 11.3 SY syllable duration modelling

The results of the experiment reported above show that the tone of the preceding and following syllables, in terms of the peak of stylised  $f_0$  curve, do not affect the duration of the target syllable. Therefore, we considered only the first seven amongst the factors listed in Section 11.1. The duration of a syllable spoken in isolation differs from its duration when it occurs in the context of an utterance. This implies that the factors that affect the duration of syllables in the context of fluent speech actually modify the canonical duration.

This modification can produce three effects on the duration of the canonical syllable: (i) decrease the duration, i.e. compresses the syllable, (ii) leaves the duration unchanged, or (iii) increases the duration, i.e. stretches the syllable. Therefore, we constructed the model for SY duration based on the assumption that the effects of all factors can be combined multiplicatively. (*Chen et al.*, 2003) has shown that such



multiplicative models perform better than additive models. If  $\lambda_i$  is the variable that represents the contribution of factor  $i$  to the duration of a syllable, we can express the resultant effects of the factors as:

$$\eta = \prod_{i=1}^7 \lambda_i. \quad (11.1)$$

Let  $L_r$  and  $L_c$  be the realised and the canonical duration of a syllable respectively. If  $\eta$  is defined such that  $-1 < \eta \leq 1.0$ , we can formulate the equation for computing  $L_r$ , given  $L_c$ , as:

$$L_r = (1.0 + \eta)L_c \quad (11.2)$$

where  $\eta$  denotes the syllable duration modifier. In Equation 11.2,  $\eta$  acts as a scaling factor for the duration of a syllable. If  $\eta = 0$ , it implies that the realised syllable duration is the same as the canonical duration of the syllable in question. This is the case when monosyllabic words are spoken in isolation or at the beginning of a short sentence.

When  $\eta < 0$  the realised duration is reduced by the factor specified by  $\eta$ . For example, if  $\eta = -0.5$ , it implies that the realised duration is 50% shorter (compressed to half of its canonical size) than the canonical duration. Likewise, if  $\eta > 0$ , the canonical duration of the syllable is increased. Our aim is to develop a model that predicts  $\eta$  by establishing a relationship between the set of factors affecting syllable duration and the duration of a syllable in the context of an utterance.

### 11.3.1 Duration data normalisation

In most duration modelling, results are analysed using the raw means of duration measurements. We did not use this approach for two reasons. First, the raw mean could be misleading in speech data like ours where the frequency distribution of syllables is unbalanced. The imbalance in our speech database resulted from the phonotactic constraints imposed by the SY language on syllables. For example, the Vn and N syllable types occur less frequently in our text database. Also Vn, syllables rarely occur at the beginning of a word while the N type syllables never end any words. That

makes it impossible to produce all syllables in all possible contexts, hence the use of raw mean will bias the result toward syllables which occur more frequently.

An alternative is to use the z-score of feature values so that our model will predict the z-score of duration rather than the exact duration. *Chung* (2002) and *King and Clark* (2004) have suggested that predicting the z-score of duration produces better results than the raw mean duration. However, to obtain a more normal distribution that will facilitate optimal performance for using z-score, we need to first convert the raw duration data using the standard *log* function (*Chung*, 2002; *Bellegarda et al.*, 2001). *Bellegarda et al.* (2001) has shown that the duration pattern of syllables is not generally additive and the choice of the usual *log* function for the factors is unlikely to produce an optimal duration model. Therefore, we did not use the z-score in our duration modelling.

To transform our duration data, we need to ensure that the duration model gives appropriate weights to data occurring less frequently such as the ones described above. To make this possible, we need a more flexible transformation framework than those described above. Following *Bellegarda et al.* (2001), we used a transformation that maps the feature values into the interval  $[0, 1]$  as follows.

Let  $A$  and  $B$  denote the minimum and maximum values of an observed factor in the training data. Let  $D$  be the value of the feature to be transformed.  $D$  is transformed into the value  $x$  using the following equation (*Bellegarda et al.*, 2001):

$$x = \frac{D - A}{B - A} \quad (11.3)$$

The transformation is used for rendering our duration data for the development of the duration models.

## 11.4 FDT in duration modelling

As discussed in Chapter 6, the major advantage that the CART has over other data-driven methods is that the CART model is more readable and often understandable by humans. This feature is particularly important when developing a duration model for a new language as it makes it possible to iteratively evaluate and improve the

model. However, SOP models have been shown to successfully handle the data sparsity problem, which is a major weakness of CART that renders it unsuitable for our purpose. Therefore, a model that represents a half-way between CART, which is probabilistic, and SOP, which is very prescriptive, is required for our duration modelling. We have found that the fuzzy decision tree (FDT) modelling technique meets this requirement. The major strength of fuzzy logic algorithms are that they are robust and flexible and they are able to cope well with interactions of linguistic attributes. Hence, they can be easily tailored to cope with small disjuncts, which are associated with large degrees of attribute interaction (*Carvalho and Freitas, 2002*).

The motivation for selecting the fuzzy decision tree approach for our duration model is founded upon the hypothesis that proportionate relationships among confounding factors that affect duration at various phonological levels can be captured by an appropriately-designed model. Such a model must, on one hand, establish a relationship between the linguistic levels and qualitative description of duration phenomena. On the other hand, it must facilitate a transparent link between qualitative descriptions and quantitative values that is responsible for the timing of speech waveform.

The FDT is an appropriate model in this context because it does not impose any arithmetic or multiplicative restrictions (relationship), or any inherent linearity by way of empirical rules. Thus it is able to exploit a very important property of interaction between factors affecting duration, namely that i.e. these interactions are often regular in the sense that the effect of one factor do not reverse that of another (*van Santen, 1994; Campbell, 2000*).

An FDT model facilitates the computation of a more globally optimal result because it has the ability to compute the relative effects of all child nodes corresponding to duration affecting factors on the duration of a syllable, before subsequently combining and aggregating them through the defuzzification process.

#### 11.4.1 Problem formulation

We can formulate the duration modelling problem as a classification/regression problem. This is because a number of independent variables (i.e. duration affecting factors) are used to compute the duration of a syllable.

Given a set of training samples composed of observed input/output pairs that consists of  $N$  labelled examples,  $\{(\mathbf{x}_n, y_n); n = 1, 2, \dots, N\}$ , derive a general model which can be used to compute output values for any new set of input. The new input may be in the training (inside-data) set or test (out-of-data) set. In the context of duration modelling, the input variables (or attributes) are the relevant parameters describing the factors affecting the duration of the unit of utterance in focus, e.g. syllable or phone, and the output would be a numerical value specifying the actual duration or modification to each speech unit in an utterance (i.e.  $\eta$ ).

To formulate this problem as a fuzzy classification/regression problem, let  $U = \{u_j\}, j = 1, \dots, n$  represent the universe of objects that describe the factors affecting the duration of a syllable. Each of these  $n$  objects is described by a collection of attributes  $A = \{A_1, A_2, \dots, A_r\}$ . Each attribute  $A_k$  measures some important features of an object and can be limited to a set of  $m$  linguistic terms  $T(A_k) = T_1^k, T_2^k, \dots, T_m^k$ .  $T(A)$  is the domain of the attribute of  $A_k$ . Each numerical attribute  $A_k$  can be defined as a linguistic variable which takes linguistic values from  $T(A_k)$ . Each linguistic value  $T_j^k$  is also a fuzzy set defined over the range of the numerical value of the variable, i.e. its Universe of Discourse (UoD). The membership function  $\mu_{T_j^k}$  indicates the degree to which object  $u$ 's attribute  $A_k$  belongs to  $T_j^k$ . The membership of a linguistic value can be subjectively assigned or inferred by a membership function defined over its UoD.

### 11.4.2 Fuzzy decision tree design

The potential of fuzzy decision trees in improving the robustness and generalisation in classification is due to the use of fuzzy reasoning. Underlying fuzzy reasoning is the concept of a fuzzy set. A fuzzy set is represented by a membership function which maps numerical data onto the closed interval  $[0, 1]$ . While in classical logic, the result of the operations of conjunction and implication are unique, in fuzzy logic there is an infinite number of possibilities. When a crisp number from the universe of the class variable is sought and the number of other restrictions on fuzzy sets and operators are applied, rules can be evaluated individually and then combined. This approach, called local inference, gives a compositionally simple alternative, with good approximation characteristics, even when all the necessary conditions are not satisfied.

## CHAPTER 11. DURATION MODELLING

A fuzzy decision tree gives results within the closed interval  $[0,1]$ , as the possibility degree of an object matching a class. Fuzzy decision trees therefore provide a more robust way to avoid misclassification. Each path of a fuzzy decision tree, from the root to a leaf, forms a decision rule, which can be represented in the form: IF( $x_1$  IS  $A_1$ ) AND ( $x_2$  IS  $A_2$ ) ... AND ( $x_n$  IS  $A_n$ ) THEN (class =  $C_j$ ). In the case of our model, each  $x_i$  represents a factor affecting duration and the  $A_i$  are the constants defined over the universe of discourse of the factor. The  $C_i$  is the duration scaling class (i.e. Increase or Decrease).

Our aim is to exploit the Fuzzy ID3 algorithm developed by *Janikow* (1998) for addressing the problem of duration modelling in the context of SY prosody modelling. The Fuzzy ID3 which we have adopted (*Janikow*, 1998) differs from the traditional ID3 algorithms (e.g. *Quinlan* (1986)) in that the algorithm does not only create a leaf node if all data belong to the same class, but it also does so in the following cases: (i) if the proportion of a data set of a class  $C_k$  is greater than or equal to a threshold, (ii) if the number of elements in a data set is less than a threshold, or (iii) if there are no more attributes for classification. More than one class name may be assigned to one leaf node. In addition to these, the fuzzy set of all attributes are defined depending on the pattern of the data. Each attribute is processed as a linguistic variable using fuzzy restrictions such as  $X_1$  IS Low,  $X_1$  IS Medium, etc. of our FDT duration model. The implementation follows the steps in the literature (e.g. *Yuan and Shaw* (1995); *Olaru and Wehenkel* (2003)):

1. Fuzzification of the training data
2. Building a set of fuzzy decision trees
3. Obtaining an optimal tree using pruning technique
4. Applying the FDT for predicting duration

In the following subsections, we describe how these steps were applied in the design and implementation of our FDT based duration model.

### 11.4.3 Fuzzification of the input space

The FDT is an approximation structure that computes the degree of membership of the factors affecting duration to a particular syllable duration scaling class (i.e. Increase

Table 11.4: Syllable duration affecting factors

No.	Affecting Factor	Type	Values/Fuzzy terms
1.	Length of word in which the syllable occurs	Numerical	Short, Medium, Long
2.	Position of the syllable in the word	Numerical	Initial, Medial, Final
3.	Position of the word in the sentence	Numerical	Initial, Medial, Final
4.	Value of $f_0$ peak of syllable	Numerical	Low, Mid, High
5.	Structure of preceding syllable	Categorical	CV, V, CV <sub>n</sub> , V <sub>n</sub> , N
6.	Structure of target syllable	Categorical	Blank/pause, CV, V, CV <sub>n</sub> , V <sub>n</sub> , N
7.	Structure of following syllable	Categorical	Blank/pause, CV, V, CV <sub>n</sub> , V <sub>n</sub> , N

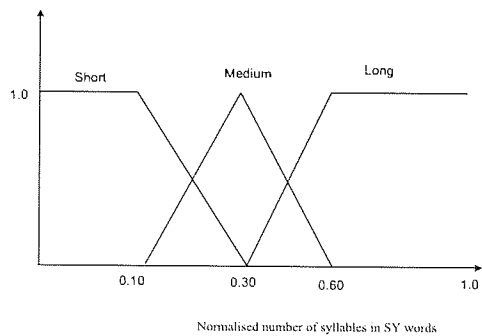
Table 11.5: Syllable duration predicted

No.	Predicted output	Fuzzy restrictions
1.	Degree to which the syllable is stretched or compressed	Increase, Decrease

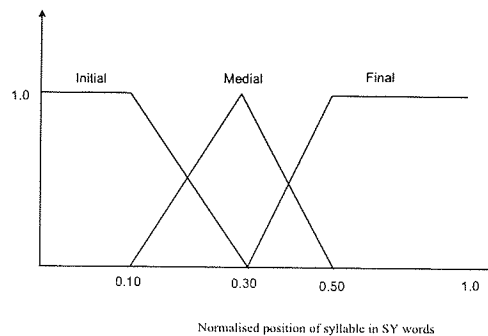
or Decrease). There are two types of data in our duration model: categorical and numerical. As shown in Table 11.4, four of the seven input variables in our duration model are numerical and are treated as continuous variables. The numerical data must be fuzzified into linguistic terms through the fuzzification process. The fuzzy membership functions used to fuzzify the numerical data are derived as follows.

We assume that these variables are factorable such that fuzzy subsets can be defined over their Universe of Discourse (UoD). Since all the factors are normalised, their UoD is defined over the closed interval  $[0,1]$ . We first partition the UoD for each of the numerical variables into subranges, with each subrange labelled with a linguistic term. For simplicity, we restrict the number of linguistic terms to 3 for continuous input variables and 2 for the output variables (i.e. Increase and Decrease). We used the trapezoidal function to model our membership functions because it is simple and there are algorithms for deriving and implementing them (Kosko, 1994). In addition, the trapezoidal membership function is frequently used in fuzzy theory to model relatively stable data such as syllable duration. The algorithm for generating the membership functions in our model is described as follows.

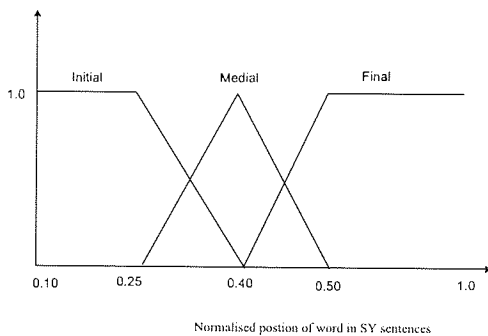
Assume that the factor  $A$  has numerical value  $x$  as computed by Equation 11.3.



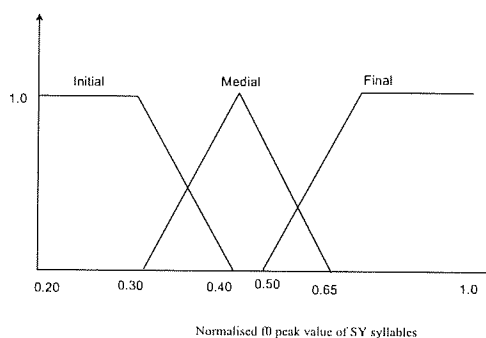
(a) Membership function for word length



(b) Membership function for position of syllable in word



(c) Membership function for position of word in sentence



(d) Membership function for peak  $f_0$  values of tone

Figure 11.3: Membership function of continuous duration affecting factors

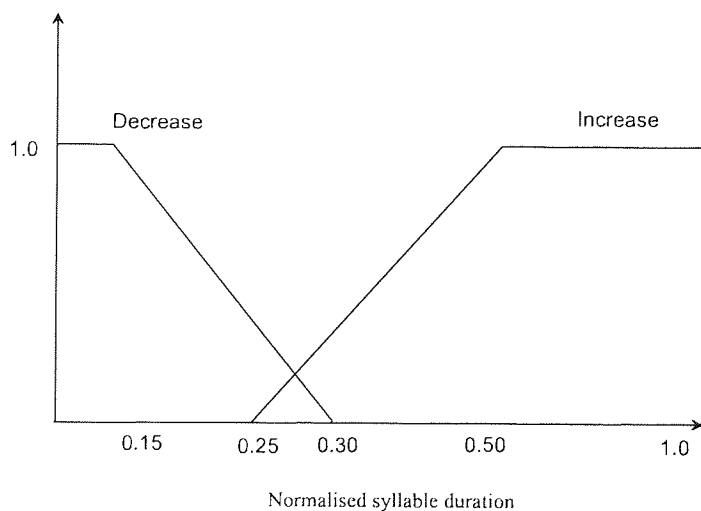


Figure 11.4: Membership function for the output

The numerical value of attribute A for all linguistic terms  $u \in U$  can be represented by  $\mu = X = \{x(u), u \in U\}$ . We defined the trapezoidal function for each variable as a four-tuple (Kosko, 1994; Mitaim and Kosko, 2001)  $(l_j, ml_j, mr_j, r_j)$  where  $ml_j \leq mr_j \in \mathbb{R}$ . The variables  $l_j > 0$  and  $r_j > 0$  denote the distance of the support of a function to the left and right of  $ml_j$  and  $mr_j$ , the centre of which is  $m_j = 1/2(ml_j + mr_j)$ . The degree to which a crisp value  $x$  belongs to the fuzzy set  $u_j$ , i.e.  $\mu_{u_j}(x) \in [0, 1]$ , is computed using the membership function:

$$\mu_{u_j}(x) = \begin{cases} 1.0 - \frac{ml_j - x}{l_j} & \text{if } ml_j - l_j \leq x \leq ml_j \\ 1.0 & \text{if } ml_j \leq x \leq mr_j \\ 1.0 - \frac{x - mr_j}{r_j} & \text{if } mr_j < x \leq mr_j + r_j \\ 0.0 & \text{otherwise} \end{cases} \quad (11.4)$$

The membership functions of each of the four input variables are shown in Figure 11.3. Figure 11.4 depicts the membership function of the output variable described in Table 11.5. It will be observed that the membership function for the output variable centers around 0.25. This is so because the distribution is skewed to low numbers and not normal distribution.

#### 11.4.4 Building the FDT

We used the FID33 software developed by Janikow (2004) for building the FDT. The variables and parameters required for implementing our FDT-based duration model are defined in Table 11.6. A speech database of 60 SY statement sentences (cf. Chapter 8) was used to generate a 250 data set. The data set is split into two disjoint parts: 220 of which formed the training set (RS) and the remaining 30 formed the testing set (TS). The training set is then divided into two other disjoint set comprising 200 growing set (GT), which is used to build the FDT, and 20 pruning set (PS), which is used to optimise the tree. The algorithm depicted in Figure 11.5 implements the tree building process.

A number of trees were generated by varying the parameters for running the FID33 program. The fuzzy decision tree shown in Figure 11.6 illustrates the structure a typical tree. Each non-terminal node of the FDT contains: (i) the attribute used to split the node (i.e. Attr), and (ii) the total example count for each decision (i.e. increase and decrease) in the node. The two values in the terminal nodes indicate the example



Table 11.6: A summary of our FDT variables, functions and parameters

Variable/function/ parameter	Description
$V_i$	A variable to represent one, i.e. the $i^{th}$ , of the duration affecting factors
$V_p^i$	A fuzzy term $p$ defined for variable $V_i$ (e.g. $V_{Short}^{Word\_Length_i}$ )
$\mu()$	The membership function $\mu_{v_i}(x)$ for variable $V_i$ defined over the crisp input $u$ . It determines how the crisp value for variable $V_i$ satisfies the restriction [ $V_i$ is $v_j^i$ ]. E.g. $\mu_{Long}^{Word\_Length}(x)$ determines the degree to which the value $x$ satisfies the fuzzy restriction [ $Word\_Length$ IS $Long$ ]. The derivation of the membership functions is explained in Section 11.4.3.
$f_1()$	An aggregation function that combines the level of satisfaction of the fuzzy restrictions of the conjunctive antecedent
$f_2()$	A function that propagates the satisfaction of the antecedent to the consequence
$X_j^N$	The membership of examples $e_j$ in the node $N$ . It is computed incrementally using $f_0$ and $f_1$ .
$X^N$	$\{X^N\}$ is the set of memberships in node $N$ for all training examples
$D_i$	Fuzzy set for the input variable $V_i$ . E.g. $D_i = \{Short, Medium, Long\}$ for $V_i = Word\_Length$
$ D_i $	Cardinality of fuzzy set $D_i$ , i.e. the number of linguistic terms defined over $V_i$ . For all of our input variables, $ D_i  = 3$
$P_K^N$	Example count for decision $V_k^c \in D_c$ in node $N$
$P^N (u^i unknown)$	The total count of examples in node $N$ with unknown values for $V_i$
$P^N V_p^i$	The total count of examples in node $N$ with $V_i = V_p^i$
$I^N V_p^i$	The information contents in node $N$ with $V_i = V_p^i$
$V^N$	The set of attribute appearing on the path leading to node $N$
$G_i^N$	Information gain computed as $I^N - I^{S_{V_i}^N}$

**Begin**

Initialisation

Starts with the entire training data set represented as  
 $E = \{e_j | e_j = (u_j^1, \dots, u_j^n, y_j)\}$  in the root node.

Initialise all weights to unity, i.e.  $W = w_k = 1.0; k = 1, 2, \dots, N$ .

**do** {

Determine Split node by:

Computing the example count for node  $N$ ,  $P^N$ , using:

$$P_k^N = \sum_{j=1}^{|E|} f_2(X_j^N, \mu_{v_k^c}(y_j));$$

$$P^N = \sum_{k=1}^{|D_c|} P_k^N$$

Computing the standard information content for node  $N$ ,  $I^N$ , using:

$$I^N = \sum_{k=1}^{|D_c|} \left( \frac{P_k^N}{P^N} \times \log \left( \frac{P_k^N}{P^N} \right) \right)$$

Calculate weighted information content by:

Searching the remaining  $I^{S_{V_i}^N}$ , for duration affecting factor  $V_i$ ,  
adjusted for missing features using:

$$I^{S_{V_i}^N} = \frac{P^N - P^{N|u^i \text{ unknown}}}{P^N} \times \frac{1.0}{\sum_{V_p^i \in D_i} P^{N|V_p^i}} \times \sum_{V_p^i \in D_i} P^{N|V_p^i} \times I^{N|v_p^i}$$

Compute information gain  $G_i^N$  for each duration affecting factor using:

$$G_i^N = I^N - I^{S_{V_i}^N}$$

Select the best duration affecting factor with the highest  $G_i^N$  value

Split the selected node  $N$  into  $|D_i|$  subnodes.  
(i.e The child  $N|V_p^i$  gets examples defined by  
 $X^{N|V_p^i} = f_1(f_0(e_j, V_p^i), X_j^N)$ )

**while** ((remaining examples with  $X_j^N < 0.0$  have a unique classification)  
**or** ( $V^N \neq V$ ))

Prune the generated trees using the algorithm in Section 11.4.5

**End**

Figure 11.5: FDT building algorithm

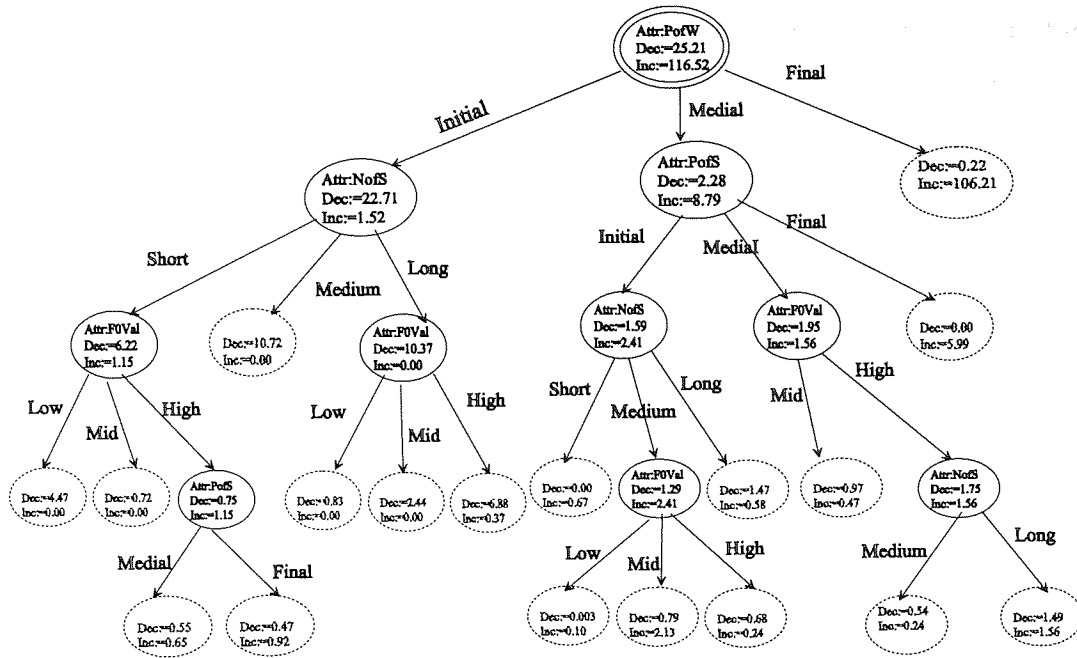


Figure 11.6: FDT for numerical values duration factor

counts for each of the two possible decisions. The example count  $N_j^N$  is computed as the membership of example  $e_j$  in  $N$ . It implies the membership in the multidimensional fuzzy set defined by the fuzzy restrictions found in  $F^N$ . It is computed incrementally using the functions  $\mu()$  (cf. Equation 11.4) and  $f_1()$  as explained in Table 11.6.

As shown in Figure 11.6, the position of a word in sentence,  $W_{tag}^{PoS}$ , produced the highest information gain over the entire data set since it is at the root of the FDT. The path along the *Final* linguistic term defined over,  $W_{tag}^{PoS}$ , leads directly to a terminal node whose example count for decreasing the syllable duration (i.e.  $Dec = 0.22$ ) is far less than that for increasing the syllable duration (i.e.  $Inc = 106.21$ ). This shows that the duration of a syllable in the final position of a sentence will increase to a very high degree. The exact amount of the increase that the syllable duration will undergo is computed by the defuzzification process explained in Section 11.4.6. Note that the tree in Figure 11.6 is built using only the duration affecting factors with numerical values.

### 11.4.5 Fuzzy decision tree pruning

Since the FDT33 program does not incorporate a pruning algorithm, the decision trees generated above vary in size and structure. This influences the performance of both

the tree and the fuzzy rules that will be extracted from it. There is a need to prune the decision tree generated in order to achieve optimal performance. In order to evaluate the efficiency of the decision trees, we applied the *T-measure* developed by Mitra *et al.* (2002). The criteria underlying the *T-measure* are:

1. The shallower the depth of the tree, the better it is since it will take less time to reach a decision.
2. The presence of an unresolved terminal node is undesirable.
3. The distribution of labelled leaf nodes at different depths affects the performance of the tree. A tree whose frequently-accessed leaf nodes are at shallower depths is more efficient in terms of time.

The *T-measure* for a decision tree is computed using Equation 11.5.

$$T = \frac{2n - \sum_{i=1}^{N_{nodes}} w_i d_i}{2n - 1} \quad (11.5)$$

$$w_i = \begin{cases} \frac{N_i}{N} & \text{for a resolved leaf node} \\ \frac{2N_i}{N} & \text{otherwise} \end{cases} \quad (11.6)$$

where  $n = 7$  is the number of attributes of a pattern,  $d_i$  is the depth of a leaf node,  $N_{nodes}$  is the number of terminal (leaf/unresolved) nodes,  $N = 200$  is the total number of patterns in the training set and  $N_i$  is the total number of training patterns that percolate down to the  $i^{th}$  leaf node. The value of  $T$  lies in the interval  $[0, 1)$ . A value of 0 for  $T$  is undesirable and a value close to 1 signifies a good decision tree. Using this measure, we select the best decision tree among those generated by the FID33 algorithm. Figure 11.7 shows the resulting tree.

The tree depicted in Figure 11.7 shows that the position of the word in which the syllable occurs in a sentence is a root affecting factor, meaning that this factor has the greatest effect on duration. When the value of the position of the word is *initial*, the decrease output decision has information gain of 65.50. The information gain for the increase output decision is 0.05. This indicates that the duration of a syllable in a word which is in the initial position has a very high degree of decrease in this situation.

If however the value of the position of word in sentence is *final*, the information gain for the decrease output decision is 0.64 while that for increase is 70.46. This indicates that the duration of a syllable in a word which is in the final position has a very high degree of increase in this situation.

```

---TREE : T-norm is Product
[] : IN=1.00 PN=150.29 : Decre=70.69 Ince=79.60
  [PositionOfWord=Initial] : IN=0.01 PN=69.61 : Decre=69.57 Ince=0.05
  [PositionOfWord=Medial] : IN=0.89 PN=13.23 : Decre=4.11 Ince=9.12
    [PositionOfWord=Medial] [FOValue=Low] : IN=0.00 PN=2.20 : Decre=2.20 Ince=0.00
    [PositionOfWord=Medial] [FOValue=Mid] : IN=0.00 PN=0.58 : Decre=0.58 Ince=0.00
    [PositionOfWord=Medial] [FOValue=High] : IN=0.40 PN=9.89 : Decre=0.77 Ince=9.12
      [PositionOfWord=Medial] [FOValue=High] [TSylType=CVn] : IN=0.68 PN=0.97 : Decre=0.17 Ince=0.80
      [PositionOfWord=Medial] [FOValue=High] [TSylType=CV] : IN=0.00 PN=3.62 : Decre=0.00 Ince=3.62
      [PositionOfWord=Medial] [FOValue=High] [TSylType=V] : IN=0.00 PN=1.00 : Decre=0.00 Ince=1.00
      [PositionOfWord=Medial] [FOValue=High] [TSylType=N] : IN=0.71 PN=2.24 : Decre=0.44 Ince=1.80
      [PositionOfWord=Medial] [FOValue=High] [TSylType=N] [PSylType=CVn] : IN=0.00 PN=0.30 : Decre=0.00 Ince=0.30
      [PositionOfWord=Medial] [FOValue=High] [TSylType=N] [PSylType=CV] : IN=0.44 PN=1.63 : Decre=0.15 Ince=1.48
      [PositionOfWord=Medial] [FOValue=High] [TSylType=N] [PSylType=N] [PositionOfSyllabe=Medial] : IN=0.87 PN=0.51 : Decre=0.15 Ince=0.36
      [PositionOfWord=Medial] [FOValue=High] [TSylType=N] [PSylType=N] [PositionOfSyllabe=Final] : IN=0.00 PN=1.37 : Decre=0.00 Ince=1.37
        [PositionOfWord=Medial] [FOValue=High] [TSylType=N] [PSylType=V] : IN=0.39 PN=0.32 : Decre=0.29 Ince=0.02
        [PositionOfWord=Medial] [FOValue=High] [TSylType=Vn] : IN=0.39 PN=2.06 : Decre=0.16 Ince=1.90
        [PositionOfWord=Medial] [FOValue=High] [TSylType=Vn] [PSylType=CV] : IN=0.87 PN=0.55 : Decre=0.16 Ince=0.39
        [PositionOfWord=Medial] [FOValue=High] [TSylType=Vn] [PSylType=V] : IN=0.00 PN=1.24 : Decre=0.00 Ince=1.24
        [PositionOfWord=Medial] [FOValue=High] [TSylType=Vn] [PSylType=Vn] : IN=0.00 PN=0.28 : Decre=0.00 Ince=0.28
  [PositionOfWord=Final] : IN=0.07 PN=71.10 : Decre=0.64 Ince=70.46
-----
TOTAL 14 leaves
-----

```

Figure 11.7: Fuzzy decision tree for the duration model

The rest of the nodes for which nodes are represented in the tree, in order of importance, are: the syllable tone, the phonetic structure of the syllable, the position of syllable in word. The information represented imitates the general pattern provided in respect of duration found in the literature (e.g. *Chen et al. (2003)*).

It should be noted, however, that the results reported here may be bias due to the type of the sentences used in the duration modelling. All the sentences starts with pronounces and it may be that syllables in pronouns are shorter. therefore, a bigger speech database will ne needed to confirm the result reported.

#### 11.4.6 Applying FDT to duration modelling

The solution provided by FDT is based on estimates made at all leaf nodes of the tree. The final decision is obtained by a collection of alternative decision paths that branch out from the root node and end at a leaf node of the FDT (*Suárez and Lutsko, 1999*). Let  $T$  denote the set of leaf or terminal nodes and let the cardinality (i.e. the number of terminal nodes) be denoted by  $|T|$ . In terms of this parameter, the total number of nodes in a tree is  $2 \times |T| - 1$  and the number of inner nodes is  $|T| - 1$ . The set  $T$  is used for actual prediction.

To compute the duration of a sample syllable,  $e_s$ , the FDT algorithm evaluates the succession of tests from the root node, following a path that is determined by the result of those tests at each internal node. Eventually this path leads to one terminal node, say  $t_i$ . The degree  $F_i^N$ , which the duration sample  $e_s$  belongs to the leaf node  $t_i$ , are then computed. The  $F_i^N$  values for all paths that starts from the root to a leaf node is computed in this manner. The final prediction is made by combining or aggregating these values. For any given vector of syllable duration factor data, the value of the predicted duration modifier is equal to the weighted average of the  $F_i^N$  values given by each of the leaves. The weight of a given leaf in the average is the degree of membership of the example to the leaf in question. The computation of the final duration modification factor is achieved by the defuzzification process.

FID33 provides a number of defuzzification schemes for achieving this goal. They included: (i) the best majority class, (ii) centre of gravity, (iii) maximum majority class. We adopt the best majority class scheme in our model because it produces

Table 11.7: CART input Description file

```

((PSylType CVn CV V N Vn Bla*)
(TSylType CVn CV V N Vn)
(FSylType CVn CV V N Vn Bla*)
(NumberOfSyllable float)
(PositionOfSyllable float)
(PositionOfWord float)
(F0Value float)
(DegOfIncren float))

```

\* Bla denotes a pause or blank.

better accuracy.

## 11.5 CART duration model

In order to compare the performance of the FDT-based duration model with the standard CART method, we implement a CART-based duration model using the Edinburgh University CART building software “Wagon” (*Black et al.*, 1999). The development of a duration model based on CART involves building a tree by training it on the input(i.e. affecting factors)-output(i.e. syllable duration modifier) data collected in respect of speech duration. The tree building algorithm successively divides the feature space to minimise the prediction error in duration values. After the tree construction phase, a relatively large tree  $T_{max}$  is obtained. Some branches on  $T_{max}$  are successively pruned resulting in a sequence of trees. The best among these trees is selected using a test sample that is independent of the training sample. This results in generating a tree which produces an optimal performance. The pruning process is done automatically.

Our CART model was built using the same set of training data and test data as in the development of the FDT. The sets are presented in the form  $\{(\mathbf{x}_n, y_n); n = 1, 2, \dots, N\}$ , where  $x_n$  are feature vectors of the corresponding affecting vector and  $y_n$  are the scaling values for syllable duration. The input file format is shown in Table 11.7.

The variables *PSylType*, *TSylType* and *FSylType* correspond to the structure of the preceding, target and following syllables respectively. The variables *NumberOfSyllable*, *PositionOfWord*, and *PositionOfWord* are the number of syllables in the word (word length), the position of the syllable in the word, and the position of the word in the

sentence respectively. *F0Value* is the peak of the stylised  $f_0$  curve of the target syllable. The variable *DegOfIncren* is the dependent variable and it is the degree of stretch or compression of the target syllable.

The tree building process starts with the tree consisting of only the root node  $t_1$  containing all cases. The task is to find the optimal binary split of the data. For real value features,  $i$ , all splits of the form  $x_i^n < \tau$  are tested, where  $\tau$  denotes a predefined threshold value. For the M-value categorical feature  $i$ , the splits have the form  $x_i \in \theta$ , where  $\theta$  goes through all subsets of the set of all possible values of features  $i$ . The best split across all features is selected and the data in the root node is split into left and right nodes, i.e.  $(t_L, t_R)$ . This procedure is applied recursively to all descendants until a stopping condition is fulfilled.

We expect our CART-based duration model to predict the value of the scale factors for the duration of a syllable which is then used to compute the realised duration. The syllable duration is calculated by the formula:

$$Duration_r = Duration_c + (Duration_r \times PrecdScale) \quad (11.7)$$

where  $Duration_c$  and  $Duration_r$  are the canonical and realised duration respectively.  $PrecdScale$  is the predicted scaling factor for compressing/stretching the syllable duration. For example, if  $PrecdScale$  is  $-0.25$ , the syllable is reduced by 25% of its original duration. If  $PrecdScale$  equals 1.0, the syllable duration is doubled (cf. Equation 11.2). A typical tree generated for numerical duration affecting factors by the CART is shown in Table 11.8. The optimal tree generated by the CART algorithm, comprising all the duration affecting factors, is shown in Table 11.9.

In the optimal CART tree of our duration model shown in Table 11.9, the tone, i.e. peak value of the  $f_0$  curve of the syllable, is the root node. This implies that the tone of the syllable is the most important factor that affects the duration of a syllable. Other factors that form the nodes, in order of importance, include the position of word in the sentence, the position of syllable in word and the phonetic structure of the following syllable. Other parameters such as the word length and the phonetic structure of the preceding syllable are not considered to have significant influence on the duration computed by the CART model.



Table 11.8: CART Tree for numerical duration affecting factors

<b>Root</b>
((PositonOfWord < 0.55)
((PositonOfWord < 0.34)
((PositonOfWord < 0.12)
((F0Value < 0.354)
((F0Value < 0.251)
((0.0015492 0.1685))
((0.000516399 0.1715)))
((F0Value < 0.498)
((0.0010328 0.183933))
((0.00573431 0.197412))))
(PositionOfSyllabe < 0.73)
((0.0125836 0.222))
((0.00771903 0.242625)))
((0.131715 0.3864)))
((PositonOfWord < 0.64)
((0.094956 0.621048))
((PositionOfSyllabe < 0.604)
((0.0636718 0.853793))
((0.040005 0.969)))

Table 11.9: Optimal CART for duration model

<b>Root</b>
((F0Value < 0.538)
((PositonOfWord < 0.37)
((PositionOfSyllabe < 0.18)
((((CVn 0.24) (CV 0.32) (V 0.12) (N 0.04) (Vn 0.08) (Bla 0.2) CV))
((FSylType is CVn)
((((CVn 0.5) (CV 0.2) (V 0.1) (N 0) (Vn 0) (Bla 0.2) CVn))
((((CVn 0.2) (CV 0.6) (V 0.0285714) (N 0) (Vn 0.171429) (Bla 0) CV))))
((((CVn 0.142857) (CV 0.571429) (V 0.122449) (N 0.102041) (Vn 0) (Bla 0.0612245) CV))))
((PositionOfSyllabe < 0.52)
((((CVn 0.0555556) (CV 0.805556) (V 0.111111) (N 0.0277778) (Vn 0) (Bla 0) CV))
((TSylType is CV)
((((CVn 0) (CV 0.3) (V 0.6) (N 0.1) (Vn 0) (Bla 0) V))
((((CVn 0.0571429) (CV 0.571429) (V 0.285714) (N 0) (Vn 0.0857143) (Bla 0) CV))))

A visual comparison of the FDT in Figure 11.7 and the CART tree in Table 11.9 shows that the CART model makes a more efficient use of the input variable by producing a more compact tree. However, the FDT approach produces a more comprehensible tree which captures the knowledge about duration phenomena reported in the literature.

## 11.6 Evaluation

In terms of theoretical computational complexity, the CART model should outperform our FDT model. This is because, in the worst case, i.e. with completely overlapping

subsets, the complexity of building a balanced fuzzy decision tree will be  $O(\|GS\|^2 \times \|a\|)$  where  $\|GS\|$  is the number of learning instances used for building the tree and  $\|a\|$  the number of candidate attributes. This evaluates to  $O(\|300\|^2 \times \|7\|) = O(6.3 \times 10^5)$  for our FDT model. This is significantly worse than  $O(\|GS\| \log \|GS\| \times \|a\|)$ , i.e. ( $O(\|300\| \log \|300\| \times \|7\|) = O(5.2 \times 10^3)$ ), which is the complexity of building a crisp decision tree. Also, the search for optimal dichotomy will be significantly more demanding in FDT than in the crisp discretisation procedure (*Boyen and Wehenkel, 1999; Olaru and Wehenkel, 2003*).

However, the theoretical evaluation does not necessarily correlate to practical performance. To obtain the practical performance of the models, we carried out qualitative and quantitative evaluations on both duration models. For the quantitative evaluation, we applied the Root Mean Square Error (RMSE) (*Hermes, 1998; Clark and Dusterhoff, 1999*) and the correlation of the actual versus predicted duration for the two models. The transcription accuracy and the Mean Opinion Scores (MOS) (*Sakurai et al., 2003; Donovan, 2003*) were used to evaluate the intelligibility and naturalness respectively in the qualitative evaluation.

The quantitative evaluation provides a performance index of how the model fits the data. A high correlation and low RMSE indicate a good fit. The results of the quantitative evaluation of the two duration models are shown in Table 11.10 & 11.11 for FDT and CART respectively. As shown in these tables, the CART-based model outperforms the FDT-based duration model for the training and test set when the evaluation is based on some individual syllable types.

For example, while the FDT model produced an RMSE of  $14.50ms$  and a correlation of 0.78 for the training set for CVn type syllables (see Table 11.10), the CART model produced an RMSE of  $12.50ms$  and a correlation value of 0.89. This pattern is repeated for all syllable types except the N type syllables where the FDT model (RMSE= $10.51ms$ , Corr=0.91) is better than the CART model (RMSE= $10.99ms$ , Corr=0.87).

When the overall duration database is considered, the CART model (RMSE= $13.99ms$ , Corr=0.88) outperforms the FDT (RMSE= $14.12ms$ , Corr=0.87) in training data but the FDT model (RMSE= $17.59ms$ , Corr=0.79) performs better than the CART model

(RMSE=22.15ms, Corr=0.75) on test data. We observed that the difference in this quantitative performance is consistent but relatively small.

Table 11.10: Result for quantitative evaluation for FDT

	Training data		Test data	
	RMSE (ms)	Correlation	RMSE (ms)	Correlation
CV <sub>n</sub>	14.50	0.78	18.20	0.71
CV	15.11	0.91	17.33	0.73
V <sub>n</sub>	12.33	0.88	15.66	0.69
V	17.05	0.79	21.56	0.66
N	10.51	0.91	15.22	0.80
Overall	14.12	0.87	17.59	0.79

Table 11.11: Results for quantitative evaluation for CART

	Training data		Test data	
	RMSE (ms)	Correlation	RMSE (ms)	Correlation
CV <sub>n</sub>	12.50	0.89	18.05	0.78
CV	17.65	0.87	22.50	0.65
V <sub>n</sub>	11.81	0.76	21.30	0.69
V	18.11	0.86	26.70	0.77
N	10.99	0.87	22.01	0.71
Overall	13.92	0.88	22.15	0.75

To put the results of our experiment in the context of other duration models for other languages in the literature, we compare our result with those obtained in some duration models (see Table 11.12). Our model performs favourably well. For example, Table 11.12 shows that the Regression model and the Hybrid (Statistical/Regression) model (*Chen et al.*, 2003) produced better quantitative results than ours. However, it is well-known that quantitative results need not correspond to the perceptual quality of the synthesised speech. In order to establish the practical performance of our models, we performed qualitative evaluations.

### 11.6.1 Qualitative evaluation

Thirty sentence are selected for this evaluation. Twenty of the sentences came from the training set while the remaining ten came from the test set. Three stimuli were created for each of the sentences producing ninety stimuli in total. Two of the stimuli are the re-synthesised speech sound generated using the duration data for the FDT

Table 11.12: Comparison of duration models (based on test set results)

Language	Model Type	RMSE(ms)	Corr.
American English ( <i>van Santen, 1994</i> )	SOP	–	0.90
Korean ( <i>Lee and Oh, 1999</i> )	CART	22.00	0.82
Korean ( <i>Chung, 2002</i> )	CART	25.11	0.77
Czech ( <i>Batůšek, 2002</i> )	CART	20.30	0.79
Mandarin ( <i>Chen et al., 2003</i> )	Regression	15.47	–
Mandarin ( <i>Chen et al., 2003</i> )	Hybrid statistical / regression	11.18	–
Mandarin ( <i>Lin et al., 2003</i> )	Recurrent Fuzzy Neural Network	20.16	–
Our SY FDT model	FDT	17.59	0.72
Our SY CART model	CART	22.15	0.75

and CART models. The third stimuli is the naturally recorded speech sound for the sentence. The stimuli were presented to the participants in a random order and they were not told which stimuli was natural or synthesised.

To obtain the re-synthesised stimuli, the duration tier of the natural speech sound was replaced with those computed by the model. We take care in applying the duration multiplier to the syllable nucleus alone to ensure proper alignment of syllable to the  $f_0$  contour of the synthesised utterance. We also ensure that the peaks and valleys of the syllables are moved relative to their citation form. The Pitch Synchronous Overlap (PSOLA) method was then used to synthesise the utterance (*Dutoit and Leich, 1993*) using the *Praat* speech processing software (*Boersma and Weenink, 2004*).

Table 11.13: Results for the intelligibility evaluation

Data set	Duration models	Intelligibly score	Significance
Training set	FDT	4.50 (0.63)	not Significant
	CART	3.80 (0.67)	( $p > 0.05$ )
Test set	FDT	4.10 (0.71)	not Significant
	CART	3.60 (0.58)	( $p > 0.05$ )

Standard deviations are shown in parentheses

### 11.6.2 Intelligibility evaluation

The intelligibility test follows the procedure discussed in Chapter 10. The only difference in this case is that after a stimuli is presented, the participant is asked to recite

what he/she heard instead of writing it down. This was to reduce the duration of the experiment since participants talk faster than they can write.

The results of the intelligibility tests are shown in Table 11.13. For the FDT-based duration model, a transcription accuracy of 4.50 (SD 0.63) was obtained for the training set. For the test set, the transcription accuracy is 4.10 (SD 0.71). The CART-based duration model produced a transcription accuracy of 3.60 (SD 0.58) for the test set and transcription accuracy of 3.80 (SD 0.67) for the training set. We observed that the standard deviations of intelligibility scores of the FDT based duration are smaller than those of the CART based model. This suggests that the FDT based duration model produced a more consistent prediction than the CART.

A sign test statistics (cf. Section 10.3.3) shows that the intelligibility scores for the two duration models are not significantly different ( $p > 0.05$ ). This results indicates that there is no evidence that the listeners preferred the synthetic speech produced using the FDT based model over that of the CART based model.

### 11.6.3 Naturalness evaluation

For the naturalness test, the participants were asked to rank the naturalness of the utterance using a scale of 1 to 5 as shown in Table 10.11. The results for the training set (see Table 11.14) show that the naturalness quality of the unmodified speech, with an MOS score of 5.0, is higher than that of the CART (3.4) and FDT (3.7). A sign test statistics shows that the the unmodified speech is highly preferred ( $p \leq 0.001$ ) by the listeners when compared with the speech generated using the CART based duration model.

Although the FDT has a higher MOS score, the naturalness quality of speech synthesised using the FDT based duration model is not significantly ( $p > 0.05$ ) better than that of the CART. This indicates that there is no evidence that the listeners preferred the synthetic speech generated using the FDT based duration model over that generated using the CART based model.

A similar pattern is repeated for the test set when the naturalness of the unmodified speech is compared with that of the synthesised speech produced by the two duration models (see Table 11.15). Also, the synthetic speech generated by the FDT

is significantly ( $p \leq 0.05$ ) more preferred over that of the CART. The smaller standard deviation obtained for the MOS of the FDT in the two tests confirms that duration predicted by FDT is more consistent than that predicted by the CART.

Table 11.14: Results for naturalness evaluation (Training set)

Comparison	Duration model	MOS score	Significance
Natural Vs. CART	Natural	5.0 (0.001)	Significant
	CART	3.4 (0.440)	( $p \leq 0.001$ )
Natural Vs. FDT	Natural	5.0 (0.001)	Significant
	FDT	3.7 (0.130)	( $p \leq 0.001$ )
FDT Vs. CART	FDT	3.7 (0.130)	not Significant
	CART	3.4 (0.440)	( $p > 0.05$ )

Table 11.15: Results for naturalness evaluation (Test set)

Comparison	Duration model	MOS score	Significance
Natural Vs. CART	Natural	4.9 (0.01)	Significant
	CART	3.1 (0.54)	( $p \leq 0.001$ )
Natural Vs. FDT	Natural	4.8 (0.02)	Significant
	FDT	3.4 (0.39)	( $p \leq 0.001$ )
FDT Vs. CART	FDT	3.4 (0.39)	Significant
	CART	3.1 (0.54)	( $p \leq 0.05$ )

## 11.7 Summary

We have presented a Fuzzy Decision Tree (FDT) and a Classification and Regression Tree (CART) based duration model in the context of duration modelling for SY text-to-speech synthesis. Since the duration modelling is syllable-based, we first carried out a set of exploratory analytical experiments to determine the effect of factors affecting duration of SY syllables. The results of this experiment led to the selection of seven factors which are then used to produce the duration models. The duration of citation versus contextual syllables are used to predict a scale factor by which the duration of

a citation syllable will be multiplied to reflect the perceptual quality of its contextual equivalent.

The results of qualitative and quantitative evaluations show that CART models the training data more accurately than FDT. The FDT model, however, is better in extrapolating from the training data since it produced a better accuracy for the test data set. Synthesised speech produced by the FDT duration model was also ranked better on quality than the CART model. These results confirmed the well-known fact that CART possesses a very good interpolation but poor extrapolation capabilities (*Breiman et al.*, 1984; *Barbosa and Bailly*, 1994). The good extrapolation capability of FDT makes it an ideal model for implementing duration for TTS application due to the sparseness of duration data.

We also observed that the expressiveness of FDT is better than that of CART. This is because the representation in FDT is not restricted to a set of piece-wise or discrete constant approximation. In addition, fuzzification of the input data imposes a continuity constraint at the boundaries of node splits in the FDT. This acts as a mechanism to limit the degree of over-fitting of the FDT.

## Chapter 12

# Prosody model implementation

In Chapters 10 and 11 we have presented how the intonation and duration dimensions are realised within the context of our R-Tree based prosody modelling. In order to synthesise prosody using these data, we need to integrate the computed dimensions into the corresponding R-Tree. The procedure for integrating the  $f_0$  and duration dimension into the R-Tree takes into account the need to harmonise the prosody data and align the data with syllable segments. In order to evaluate our proposed model objectively, we have developed a Stem-ML based model and compared it with our R-tree based prosody model.

### 12.1 Overview of the implementation strategy

The implementation of our prosody model involves the development of a program using the *Praat* speech processing software. This program accepts two input files: a text description file and a text R-Tree file. The text description file contains information extracted from the XML based mark-up for the text. The text R-tree file contains the R-Tree for each sentence in a text. The R-Tree file contains information about the speech prosody in terms of the  $f_0$  and duration data of each syllable in the utterance to be synthesised.

After adding the  $f_0$  and duration data to the phonological structure of an utterance in the form of an S-Tree, the generated R-Tree combines phonological information with phonetic knowledge. Basically, the synthesis process assigns phonetic parameters



to pieces of the phonological structure. Perceptual information is integrated into the computed  $f_0$  through the stylisation process. Expert knowledge is incorporated into the model through the use of fuzzy logic based rules and models. This process enables our model to mimic as much as possible the systematic spectral temporal and intonational details observable in natural speech (Ogden *et al.*, 2000).

As discussed in Chapter 7, our speech synthesis approach uses syllable overlap strategy for synthesising word and word concatenation strategy for synthesising utterance units longer than a word, e.g. a phrase or a sentence. We also assume that most of the prosodic features of an SY utterance can be described by the parameters of the syllable rhyme or the voiced portion of the syllable. In the case of duration, for example, we assume that the syllable onset, usually a consonant, remains unchanged when the duration of the syllable changes.

As an illustration, if the total duration of a CV type syllable is  $D_{total}$ , the duration of the onset is  $D_{onset}$  and the duration of the nucleus is  $D_{nucleus}$ . The total duration of the syllable will be expressed as:

$$D_{total} = \langle D_{onset}, D_{nucleus} \rangle \quad (12.1)$$

If the syllable is compressed by, say degree  $n$ ,  $0.0 \leq n \leq 1.0$ , then  $D_{nucleus}$  will be reduced by the factor  $n$  of its original canonical value as well as the factor  $n$  of the onset's canonical value. This will ensure that the total duration of the syllable is reduced by  $n$  at the compressible portion, i.e. the rhyme portion, as specified in the equation:

$$\begin{aligned} D_{compressed} &:= \langle D_{total} - (D_{total} \times n) \rangle \\ &:= \langle D_{onset}, D_{nucleus} - (D_{nucleus} \times n + D_{onset} \times n) \rangle \end{aligned}$$

The syllable duration is defined operationally in that the sign of the  $n$  value determined whether the syllable is compressed or expanded.  $n = +0.5$  will yield a syllable which is about one and a half longer than the canonical syllable in terms of duration.  $n = -0.5$  will have the opposite effect and  $n = +1.0$  doubles the duration of the syllable.

The application of this formula is, however, dependent on the phonological identity of the objects that make up the syllable. Overlapping syllables to synthesise a word

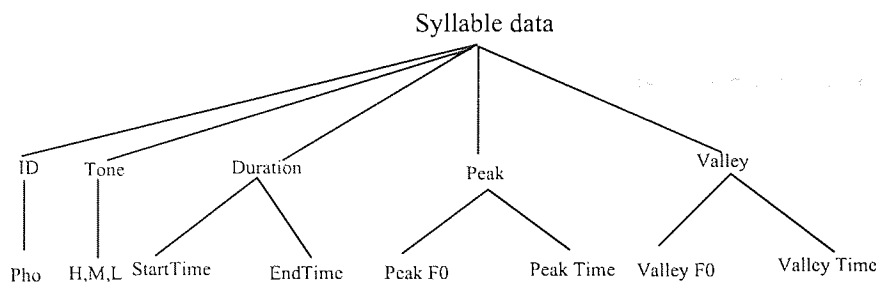


Figure 12.1: Structure of syllable data file

requires the temporal co-ordination of the events at the end of one syllable and the beginning of the next (Coleman, 1994). In addition to this, our holistic approach to prosody modelling also requires that the  $f_0$  parameters must be synchronised with the duration data and aligned with the syllable structure. We note also that the phonetic properties of components of the syllables, i.e. consonants and vowels, contribute to the complete  $f_0$  contour specification. In our model, we assume that this low level contribution is already inherent in the acoustic data of the respective syllables participating in the syllable overlaying operation.

The syllables in our database are recorded with *Wavesurfer* and annotated using *Praat*. Samples of the labelled files are shown in Appendix E. Each recorded syllable is annotated and the corresponding *TextGrid* file is generated. The *TextGrid* file contains information about the phonetic and tonal identification of the syllable as well as the duration data. In addition, the  $f_0$  data are extracted from the *Pitch tier* file and stored in a separate text file.

The  $f_0$  file contains the  $f_0$  peak and valley as well as the time of the  $f_0$  peak and valley. The phone and tone identification of each syllable are also included in the file. The information in the *TextGrid* and  $f_0$  data files are used to create the syllable data file for the speech database. The structure of the syllable data file is shown in Figure 12.1.

The phone of the syllable is the orthographic representation of the syllable when the tone is ignored. The tone of the syllable is represented by the letters *H*, *M* and *L* for high, mid and low tones respectively. For example, the phone and tone of the syllable *Bá* is coded using *Ba* and *H* respectively.

### 12.1.1 Alignment model

We consider an  $f_0$  contour representing an utterance as containing a number of  $f_0$  curves, each occupying a time slice in the  $f_0$  contour of the utterance. Two crucial issues arise in implementing our prosody model: (i) how to synchronise the  $f_0$  data with the duration data and (ii) how to align the synchronised  $f_0$  and duration data with the phonetic structure of the syllable. We have proposed an alignment model to deal with these issues.

The time alignment model considers how the stylised  $f_0$  curve is positioned with respect to time in relation to the segmental structure of each syllable in the utterance.  $f_0$  alignment considers how the  $f_0$  parameters are positioned with respect to the overall  $f_0$  contour of the utterance.

To model the  $f_0$  alignment, we need to determine tonal targets for the  $f_0$  curve of each tone type. Intuitively, a tonal target can be viewed as the maximum and minimum value within the  $f_0$  curve of each syllable (*D'Imperio*, 2002). Determining the exact tonal target for each syllable in an utterance is very difficult because the temporal alignment of  $f_0$  targets seems to be affected by a number of phonetic and phonological factors that can potentially interfere with how the  $f_0$  aligns in an utterance.

Recent studies on SY tones (*Akinlabí and Liberman*, 2000; *Láníran and Clements*, 2003) suggest that Yorùbá tonal targets are implemented along with the syllable that carries it. *Dilley et al.* (2005) have also shown that pitch accent is independently aligned with respect to the segmental string, and not with respect to each other. This implies that the alignment of tonal targets can be specified relative to segmental positions. It also implies that there is no fixed time interval between tones. However, for the canonical syllable, we can assume that the tonal target starts and ends within the duration of the syllable. We observed generally in our data that the tonal targets are restricted to the nucleus. For example, in syllables that do not contain an onset, i.e. V, Vn, or N type syllables, or those in which the onset is made up of voiced consonant, the tonal target tends to begin and end with the syllable.

These findings show that, in continuous speech context, alignment involves an abstract relationship between the peaks and valleys of the  $f_0$  curves with the phonetic segment of the syllable. This abstract relationship can be implemented using a com-

putational model which maps between intonation and segmental trajectories that is invariant within a given phonological category of the speech waveform. Therefore, the alignment of tonal targets can be specified relative to the duration of the segmental components of the syllable. The variation is determined based on the duration tier of the speech waveform. The following subsections discuss the alignment model which forms the basis of our speech synthesis.

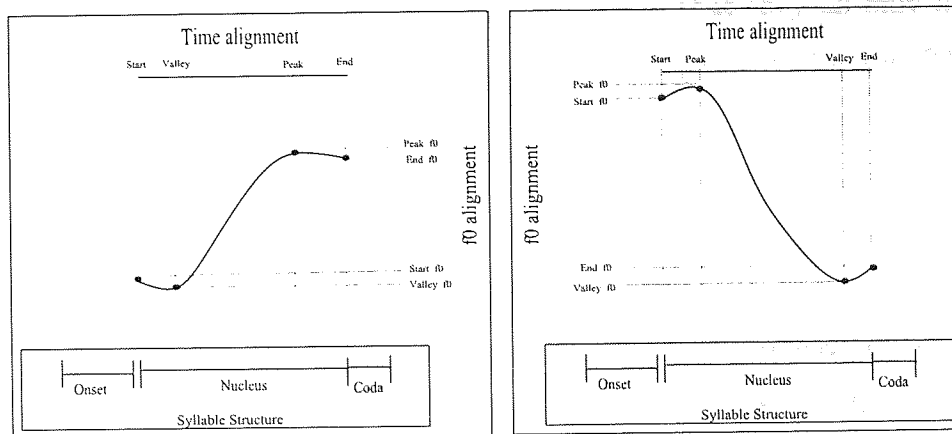
### 12.1.2 CVn type syllable alignment

The alignment model for a CVn type syllable is shown in Figure 12.2 a & b. This model assumes that the  $f_0$  for the syllable starts and ends on the vowel that forms the nucleus of the syllable. In the case of the H tone curve, the valley appears nearer to the onset and the peak appears nearer to the coda of the syllable (see Figure 12.2(a)). The M tone curve follows that of the H tone, only that the range, i.e. the difference between the peak and the valley, is much smaller. Similar to the H tone, the L tone  $f_0$  curve is positioned on the vowel that forms the nucleus of the syllable. However, the peak is closer to the onset and the valley is closer to the coda of the syllable (see Figure 12.2(b)).

The exact position, in time, of the peak and the valley varies with the phonetic component of the syllable.

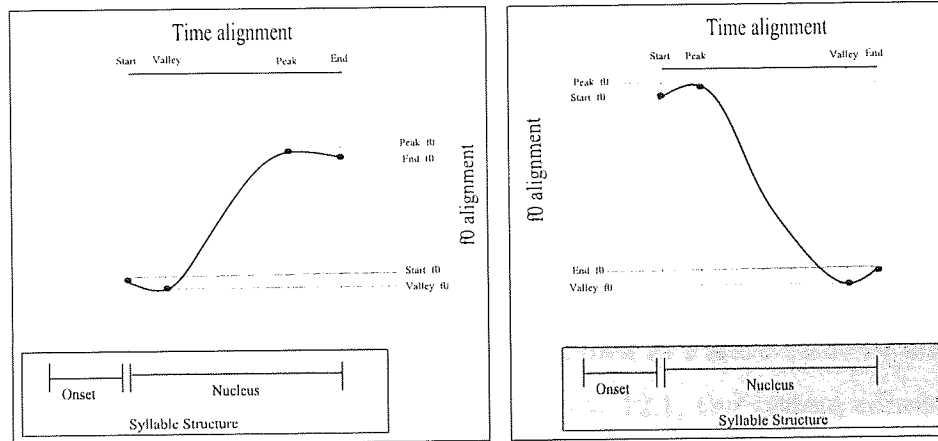
### 12.1.3 CV type syllable alignment

The model for a CV type syllable is shown in Figure 12.2 c & d. This model, like the CVn model, assumes that the  $f_0$  for the syllable starts and ends on the vowel that forms the nucleus of the syllable. In the case of the H tone, the  $f_0$  curve minimum, i.e. the valley, starts early in the onset and increases to reach its peak just before the end of the onset (see Figure 12.2(c)). Similarly, the M tone  $f_0$  curve follows that of the H tone, only that the range, i.e. the difference between the peak and the valley, is much smaller. The L tone  $f_0$  curve is also positioned on the vowel that forms the nucleus of the syllable. However, the  $f_0$  curve peaks at the beginning of the nucleus and decays to its minimum value, i.e. the valley, just before the end of the nucleus (see Figure 12.2(d)).



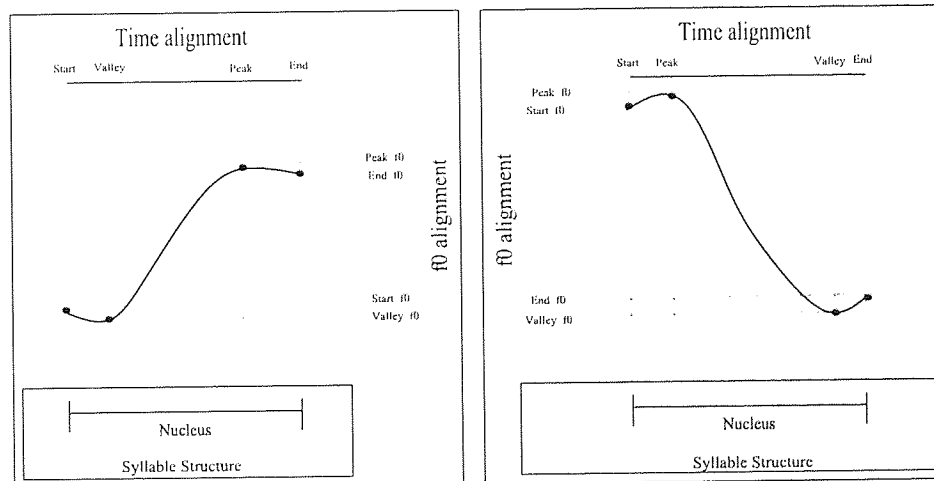
(a) H tone alignment for CVn syllable

(b) L tone alignment for CVn syllable



(c) H tone alignment for CV syllable

(d) L tone alignment for CV syllable



(e) H tone alignment for V syllable

(f) L tone alignment for V syllable

Figure 12.2: Schematics of time and  $f_0$  alignment

### 12.1.4 Other syllable type alignment

The alignment model in the V, Vn and N are similar in that they all only have the rhyme but no onset. The model for V type syllables is depicted in Figure 12.2 e & f. Since the syllable is made up of voiced component only, the  $f_0$  curve covers the entire syllable. The H tone model has its minimum  $f_0$  early in the syllable and the  $f_0$  curve reaches the maximum, toward the end of the syllable (see Figure 12.2(e)). The reverse is the case of the L tone as shown in Figure 12.2(f).

The time alignment is relative in the sense that, the time location of the start, peak, valley and end are computed relative to the canonical versus predicted duration scaling factor. If, for example, the scaling factor for a syllable is -0.15, this implies that the duration of the canonical syllable must be reduced by 15%. The new or synthesised start, peak, valley and end are then computed as 15% of the canonical *start*, *peak*, *valley* and *end*.

## 12.2 Waveform synthesis

A schematic diagram of the synthesis process is shown in Figure 12.3. In addition to the syllable parameter database discussed in Section 12.1, the system consists of five additional components. The first component is the syllable search/selection component which searches a sequence of *Praat TextGrid* files for all the syllables in the utterance to be synthesised. The output of the search/select component is used to compute each of the  $f_0$  and duration parameters of the utterance to be synthesised using the model discussed in Chapters 10 and 11. This aspect of the model is implemented using the programming languages C and MATLAB 6.5 (see Appendix F). The  $f_0$  and duration data computed are used by the syllable overlay and word concatenation processes to generate the temporal structure of the acoustic waveform. Finally, the PSOLA technique is applied to synthesise the utterance. The last two modules of the synthesis system are implemented using Praat script.

A Praat script loads the respective speech waveform files for each syllable and extracts pitch information from the sound file. To do this, a sound object is first converted to a manipulation object through the "To Manipulation" function. When a sound ob-

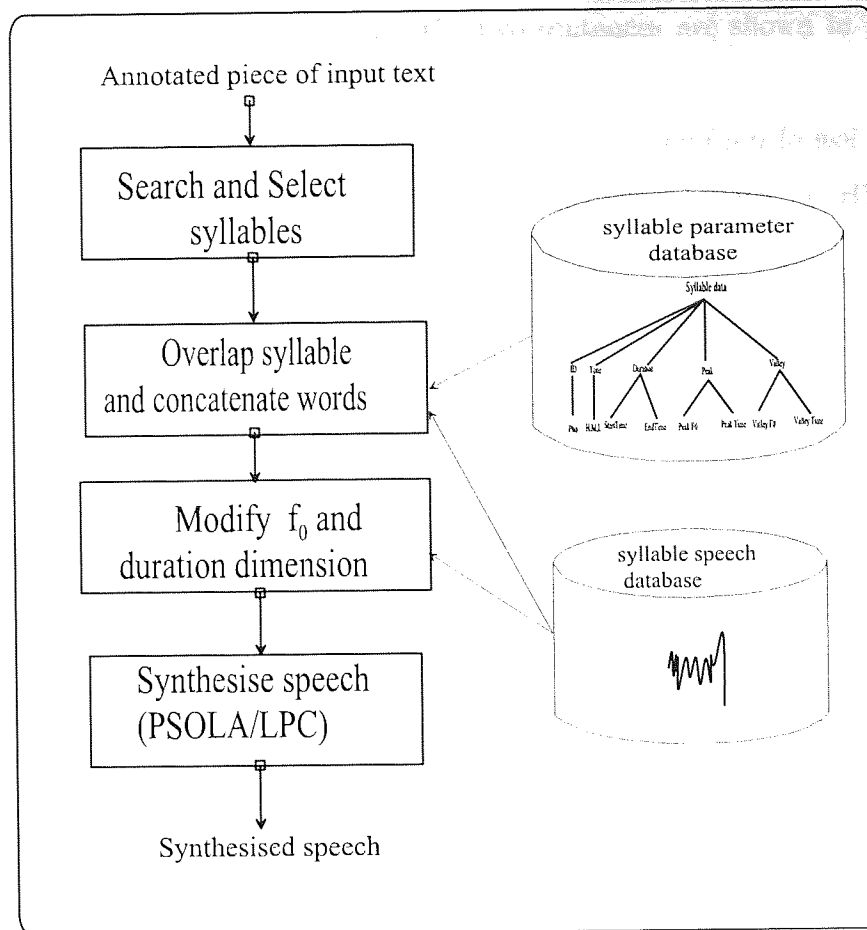


Figure 12.3: The schematics of speech synthesis process

ject is converted to a manipulation object, Praat automatically calculates the sound's  $f_0$  information using the PSOLA method. After selecting a manipulation object, the "Extract pitch tier" function is used to obtain the  $f_0$  data. The manipulation objects pitch tier is then replaced with the pitch tier object calculated by the C program. Both objects are selected and the "Replace pitch tier" function is used to apply the computed pitch parameter for synthesising the new speech sound. The duration tier is also manipulated in a similar manner.

### 12.3 Illustrations of the synthesis process

We illustrate our model with two SY sentences: (i) "Óní láti lọ wobè" (meaning "[He must go see place] He must go and see the place") and (ii) "Òdòmí lódé, kó tó lọ" (meaning "[Place me come he, before go] He came to my place before going"). The

transcripts for the S-Tree generation for the two sentences are shown in Figure 12.4 and Figure 12.5 respectively.

The abstract waveform for the S-Tree for the sentence “*Óní látí lẹ wobè.*” shows a general downward trend. This is due to the nature of tone sequence, i.e. HHHMMML. The first three tones are high tones. These tones are higher in pitch than the next three tones, which are mid tones. The mid tones are also higher in pitch than the last tone, which is a low tone. We therefore expect an  $f_0$  waveform which is generally declining with a much steeper fall at the end due to the presence of an L tone at the very end of the utterance.

The sentence “*Òdòmi lódé, Kó tó lẹ*” exhibits a more complex tonal pattern. First, it has two intonation phrases because it is made up of two phrases. The first phrase, i.e. “*Òdòmi lódé*”, should have a general upward pitch pattern based on the LLMHH tone sequence in the phrase. The second phrase, i.e. “*Kó tó lẹ*”, should have a general downward trend as a result of the HHM tone sequence in the phrase.

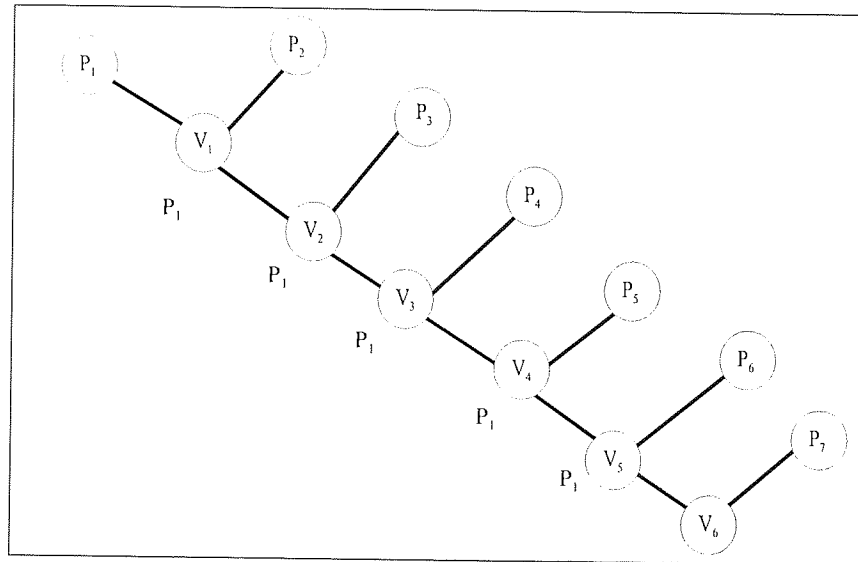
These generally expected patterns are predicted by the phonological rules as depicted by the S-Tree and abstract waveforms of the two sentences. The computed peaks, valleys and the time of peaks and valleys for the two sentences are shown in Table 12.1 and Table 12.2. These data are used to synthesise the corresponding speech as explained in Section 12.2. The corresponding synthesised utterance waveform,  $f_0$  and duration data are plotted in Figure 12.6 and Figure 12.7. In Figures 12.6 & 12.7, the dashed lines indicate the synthesised  $f_0$  contour. The continuous lines indicate the natural  $f_0$  contour. The ‘\*’ indicates sentence boundaries and the comma (,) indicates a phrase boundary.

As shown in Figures 12.6 & 12.7, the synthesised  $f_0$  contours differ from the natural one but generally follows the pattern of the natural  $f_0$  contour. The synthesised  $f_0$  also shows some blunt angles at syllable junctions. These angles result in some perceptible clicks in the synthesised speech sound. We observed that the  $f_0$  curve for syllables in the synthesis utterance does not always show peak and valley as is the case in the canonical syllables. This is particularly prevalent in the M tone syllable, where the  $f_0$  points computed as peaks and valleys usually have the same  $f_0$  values. This results in a line rather than a curve. Furthermore, some H and L tones have flat or slightly

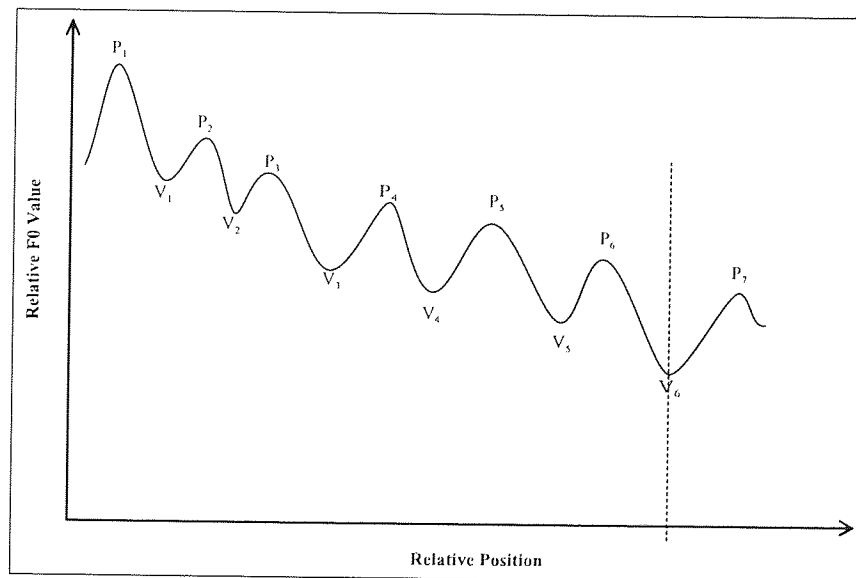


Sentence	Óní láti lọ wobè.						
Words	Óní		láti		lọ	wobè	
Syllables	Ó	ní	lá	ti	lọ	wo	bè
Tones	H	H	H	M	M	M	L
Peaks	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$	$P_7$
Peak-pairs	$(P_1, P_2)$	$(P_2, P_3)$	$(P_3, P_4)$	$(P_4, P_5)$	$(P_5, P_6)$	$(P_6, P_7)$	
Peak-valley Structures	$((P_1, P_2), V_1)$		$((P_2, P_3), V_2)$		$((P_3, P_4), V_3)$		$((P_4, P_5), V_4)$
							$((P_5, P_6), V_5)$
							$((P_6, P_7), V_6)$

(a) Association of peaks and valleys with syllables



(b) Skeletal tree



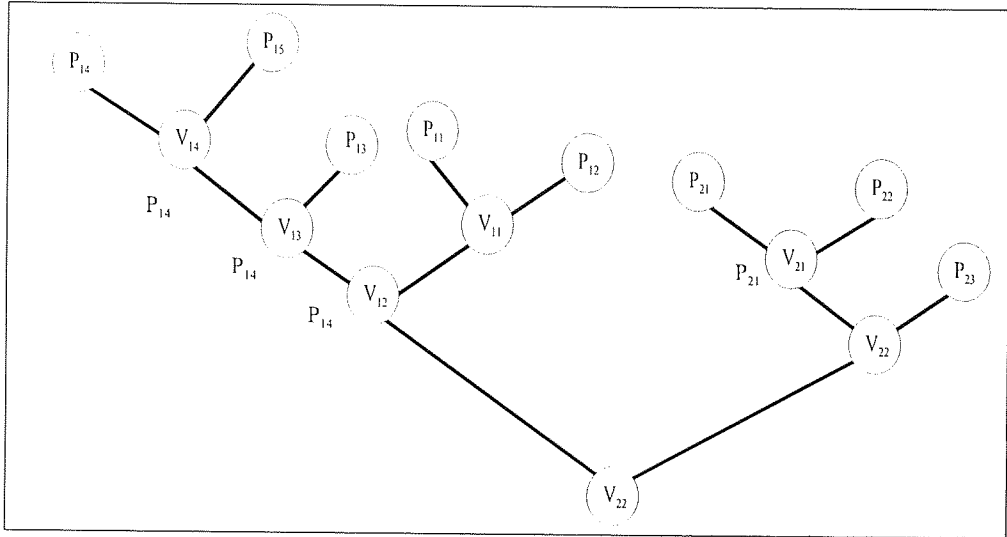
(c) Abstract waveform

Figure 12.4: Transcript of *S-Tree* generation for the sentence "Óní láti lọ wobè."

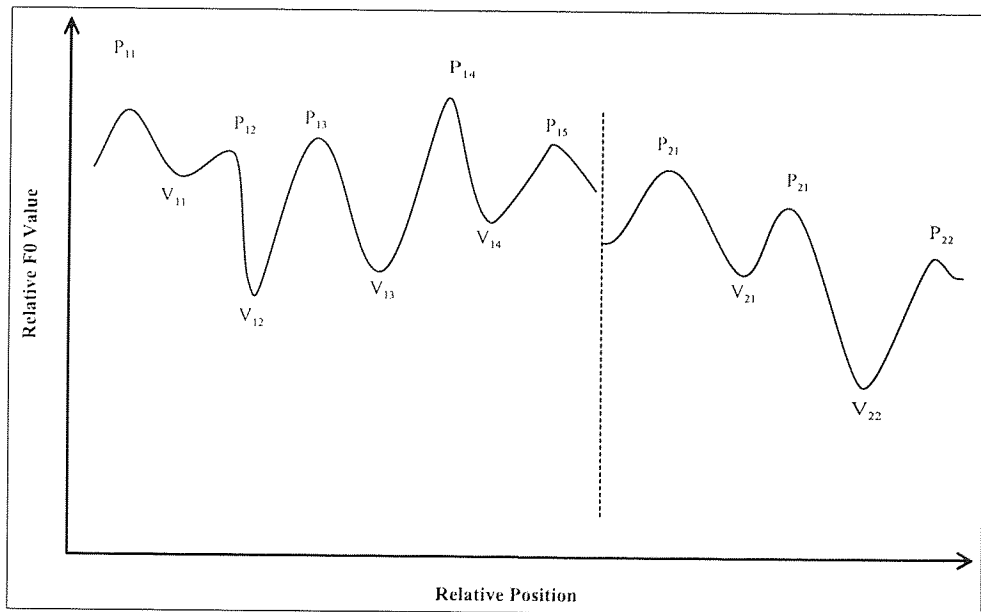
CHAPTER 12. PROSODY MODEL IMPLEMENTATION

Sentence	Òdòmi lódé, Kó tó lọ.							
Words	Òdò		mi	lódé		Kó	tó	lọ
Syllables	Ò	dò	mi	ló	dé	Kó	tó	lọ
Tones	L	L	M	H	H	H	H	M
Peaks	$P_{11}$	$P_{12}$	$P_{13}$	$P_{14}$	$P_{15}$	$P_{21}$	$P_{22}$	$P_{23}$
Peak-pairs	$(P_{11}, P_{12})$	$(P_{12}, P_{13})$	$(P_{13}, P_{14})$	$(P_{14}, P_{15})$		$(P_{21}, P_{22})$	$(P_{22}, P_{23})$	
Peak-valley Structures	$((P_{11}, P_{12}), V_{11})$	$((P_{12}, P_{13}), V_{12})$	$((P_{13}, P_{14}), V_{13})$	$((P_{14}, P_{15}), V_{14})$		$((P_{21}, P_{22}), V_{21})$	$((P_{22}, P_{23}), V_{22})$	

(a) Association of peaks and valleys with syllables



(b) Skeletal tree



(c) Abstract waveform

Figure 12.5: Transcript of *S-Tree* generation for the sentence “Òdòmi lódé, Kó tó lọ.”

Table 12.1: Computed data for the sentence “*Óní láti lọ wobè.*”

Syllable	$f_0$ peak (Hz)	Time of $f_0$ peak (sec)	$f_0$ valley (Hz)	Time of $f_0$ valley (sec)
Ó	148.000	0.126	135.900	0.051
ní	145.500	0.278	145.200	0.221
là	149.200	0.346	140.700	0.311
ti	115.600	0.561	108.300	0.638
lọ	121.700	0.690	114.400	0.781
wo	112.800	0.800	110.300	0.881
bè	104.000	1.000	77.200	1.110

Table 12.2: Computed data for the sentence “*Òdòmí lódé, Kó tó lọ.*”

Syllable	$f_0$ peak (Hz)	Time of $f_0$ peak (sec)	$f_0$ valley (Hz)	Time of $f_0$ valley (sec)
Ò	109.051	0.093	101.760	0.153
dò	112.320	0.173	105.042	0.273
mi	126.029	0.405	116.264	0.303
ló	138.915	0.523	127.731	0.463
dé	137.854	0.773	122.586	0.653
Kó	133.950	0.933	124.067	0.993
tó	138.553	1.093	123.230	1.188
lọ	113.912	1.322	111.390	1.373

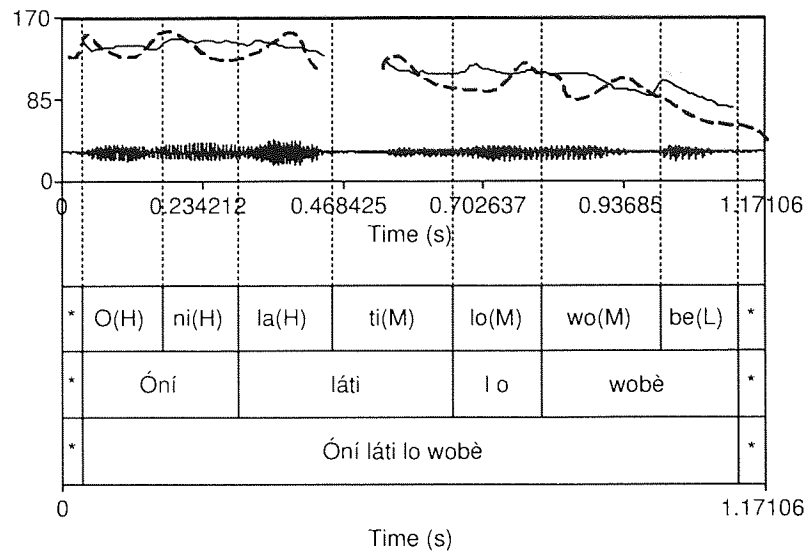


Figure 12.6: Result for the sentence “Óní láti lo wobè.”

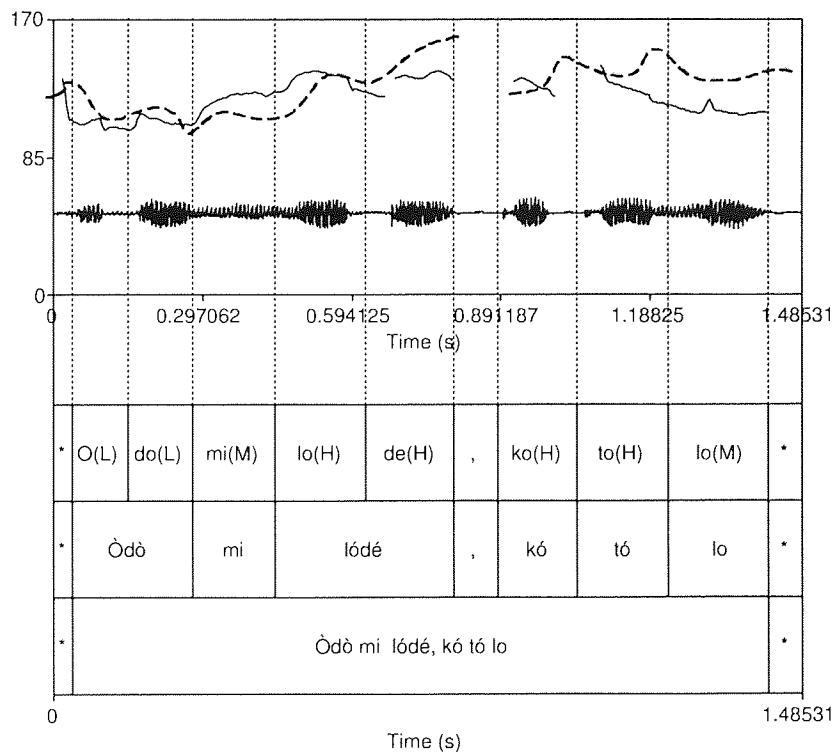


Figure 12.7: Result for the sentence “Òdò mi lódé, Kó tó lo.”

tilting  $f_0$  curves. Our observation, however, agrees with the findings of recent studies on Mandarin Chinese (Xu, 1998, 1999b), which suggest that a tone may have a peak only when it is in an appropriate tonal or prosodic context.

We observed some unexpected results in terms of the  $f_0$  peak of some H tones being lower than their valleys, e.g. the syllable dé in Figure 12.7. A speculative explanation for that may be that the  $f_0$  curves of the tones in question suffer some co-articulatory effects as well as boundary perturbations since such phenomena mostly occur at the word, phrase and sentence boundaries. However, there are no available theoretical findings in the literature to support this claim.

## 12.4 Stem-ML model for SY

In order to put the result of our work in the context of contemporary work, we also developed a simple Stem-ML model. The Stem-ML model was proposed by *Kochanski and Shih* (2003) and it has been applied to model intonation in Mandarin and Cantonese. The apparent success of Stem-ML in modelling prosody in other tone languages, i.e. Mandarin and Cantonese Chinese (*Kochanski et al.*, 2003a,b), motivated us to perform this experiment. The implementation of the Stem-ML model for SY is discussed in the following subsections. This model was developed with the help of Prof. Greg Kochanski of Oxford University Phonetic Laboratory.

### 12.4.1 Design of the Stem-ML model

The assumptions in our model, beyond those that are generic to Stem-ML (see *Kochanski and Shih* (2003)), are that:

- Each syllable carries a soft intonation target with one of the three shapes, chosen by the lexical tone. Each target is a line segment.
- The prosodic strength of a syllable both affects the precision with which a tone is realised and scales the  $f_0$  range of that tone's template. One can expect that a linguistically stronger syllable will have both a larger pitch range and also be articulated more carefully. We include an adjustable parameter in the model (*atype*) to account for such a correlation.
- Minimal syntactic information is required to model the intonation of SY. Our model includes five syntactic classes of syllables: phrase initial, phrase final, sentence initial, sentence final, and medial (i.e. everything else).

- We use no phonological rules to model intonation.
- Syllables in one- and two-syllable words have the same phonetic realisations, and we assume that there are no intrinsic differences between the first and second syllables in a two-syllable word.
- The strength of each syllable is given by:

$$S_i = A[\tau(i)] \cdot C[P(i)], \quad (12.2)$$

where  $\tau(i)$  returns the tone of the  $i^{\text{th}}$  syllable,  $A[\tau]$  is the intrinsic strength of a syllable of tone  $\tau$ ,  $P(i)$  returns the position in the sentence (e.g. sentence final, medial, . . .), and  $C[P]$  is the strength factor for position  $P$ .

- Segmental effects that depend on the phoneme sequence are relatively small.

The overall model uses 22 parameters: eleven to specify the templates, seven to specify the strengths, and four global, speaker-specific parameters. We kept the model simple to allow it to be trained on a small data set.

The model was fitted to the corpus under the assumption that the  $f_0$  data had independent Gaussian errors, using a Bayesian Markov Chain Monte Carlo algorithm that produced samples from the posterior distribution of the parameters. Once the algorithm had converged to a stationary distribution, we collected the last 6000 samples. From that, we computed the average values of all the parameters and their uncertainties.

The Stem-ML implementation is quantised in  $10ms$  increments, which raises the possibility of spurious local minima. To check for this, we started ten more Monte Carlo runs with different fixed values of the *centershift* parameter. We chose 10 random samples of parameters from the second thousand iterations from each run and computed the RMSE. The resulting set of samples traces out a minimum of RMSE against a *centershift* that is consistent with the error reported in the following chapter.

### 12.4.2 Analysis of best-fit parameters of the intonation model

The names of the Stem-ML parameter used in our model are discussed as follows. The parameter names and their values are given in bold italics.

***smooth=0.062±0.006***: This controls the rate at which the speaker changes pitch for a weak accent. It corresponds to a time of  $106±8$  milliseconds, roughly double the value obtained for a Mandarin speaker (*Kochanski and Shih, 2003*).

**base**= $105 \pm 0.5$  Hz: The speaker's base frequency.

**atype**= $0.41 \pm 0.05$ : This describes how much the pitch range of a template expands as the strength changes. It indicates a fairly weak (but significant at  $P < 0.01$ ) effect: a 40% change in strength (e.g. changing from a sentence-initial to a medial syllable) would mean that the template of the stronger syllable has an  $f_0$  range just 14% larger than a comparable medial syllable. The value is similar to the  $0.87 \pm 0.7$  value from *Kochanski et al.* (2003a).

**ctrshift** and **wscale**: These two parameters describe the scope of the template relative to the syllable boundaries. In our model, the length of the target is  $85 \pm 2\%$  of the length of a syllable, similar to the  $88 \pm 1\%$  for Mandarin. The target is nearly centered in the syllable,  $1 \pm 1\%$  of its width (i.e. about 2 ms) before the syllable centre. Combined with Stem-ML's time-symmetric mathematics, this implies a nearly equal balance between anticipatory and carry-over co-articulation.

**H tone**: The template slopes up from 8% above *base* to 112% above *base*. The target shape is both high and rising. The tone's *type*= $0.72 \pm 0.06$ . This implies that both the height and the shape are important, but errors in the tone's average value are more important than errors in the tone's shape.

**M tone**: The template slopes down from 28% above *base* to 6% above *base*. This is a mid-level and weakly falling tone. The *type*= $0.59 \pm 0.06$  parameter indicates that both the pitch height and slope are important.

**L tone**: The template slopes down from 15% below *base* to 68% below *base*. The *type* for the L tone is  $0.148 \pm 0.04$ , indicating that the shape is more important than the average value. It might be best to describe this tone as "falling" with a tendency toward low.

**Intrinsic Strengths of Tones**: For the intrinsic strength of tones, we obtain  $A[H] = 1.8 \pm 0.12$ ,  $A[M] = 2.5 \pm 0.15$ ,  $A[L] = 2.2 \pm 0.08$  (see Equation 12.2). The high tone is the weakest and the M tone is the strongest, thus there is a tendency for the shape of a high tone to be influenced by its environment than for the other tones. Although the differences are significant ( $P < 0.01$ ), it is not dramatic.

**Sentence Boundaries**: The strength factors for syllables in sentence-initial and sentence-final positions (Equation 12.2) are  $C[SI] = 1.42 \pm 0.06$  and  $C[SF] = 0.70 \pm 0.03$  respectively. Thus, with  $P < 0.01$ , sentence-initial syllables are stronger (i.e. they are articulated more precisely and have wider  $f_0$  swings) than medial syllables, and sentence-final syllables are weaker than medial syllables.

**Phrase Boundaries**: The strength factors for syllables in phrase-initial and phrase-final positions (but not at the beginning or the end of a sentence) are  $C[SI] = 1.09 \pm 0.03$  and  $C[SF] = 0.73 \pm 0.06$ . Again, with  $P < 0.01$ , phrase-initial syllables are stronger than medial syllables (though just slightly), and phrase-final syllables are also weaker than medial syllables.

The results for sentence- and phrase-boundaries parallel the results from *Kochanski and Shih* (2003) for Mandarin and *Lee et al.* (2002a) for Cantonese. The

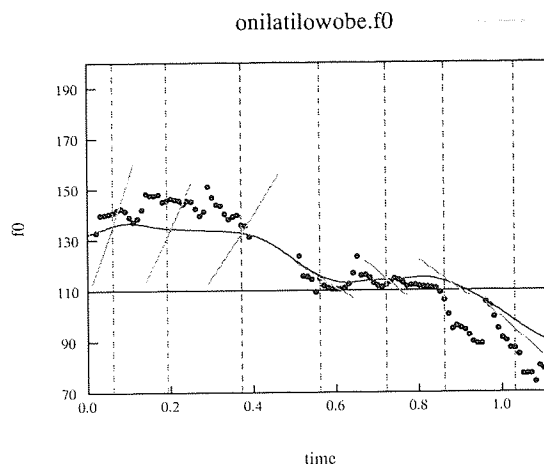


Figure 12.8: Model fit and raw data for SY sentence “*Óní láti lọ wobè*”.

differences in strength may be cues that help the listener to find phrase and sentence boundaries.

The result of applying the Stem-ML model to SY prosody modelling on our earlier sentences sample: “*Óní láti lọ wobè*” and “*Òdòmí lódé, Kó tó lọ.*” are shown in Figure 12.8 and 12.9 respectively. In these figures, the black dots mark measured  $f_0$ , the grey curve is the predicted  $f_0$ , the grey lines show the Stem-ML templates, and the vertical dashed lines show the syllable centers.

In Chapter 13, we shall discuss the results of the evaluation of the Stem-ML model and our R-tree based model.

## 12.5 Discussion

The implementation process of Stem-ML is less complex when compared with the R-Tree approach. The complexity in implementing the R-Tree based approach arises from the incorporation of perceptual data into the R-Tree approach, using the stylisation techniques. In addition, there is the need to develop a separate intonation and duration models for incorporation into the R-Tree. Also, the generation of the S-Tree using phonological rules presupposes that all intonation phenomena can be predicted using phonological rules.

Our R-Tree based prosody model derives its computational power from phonological and phonetic knowledge, unlike the Stem-ML which is a quantitative model based on



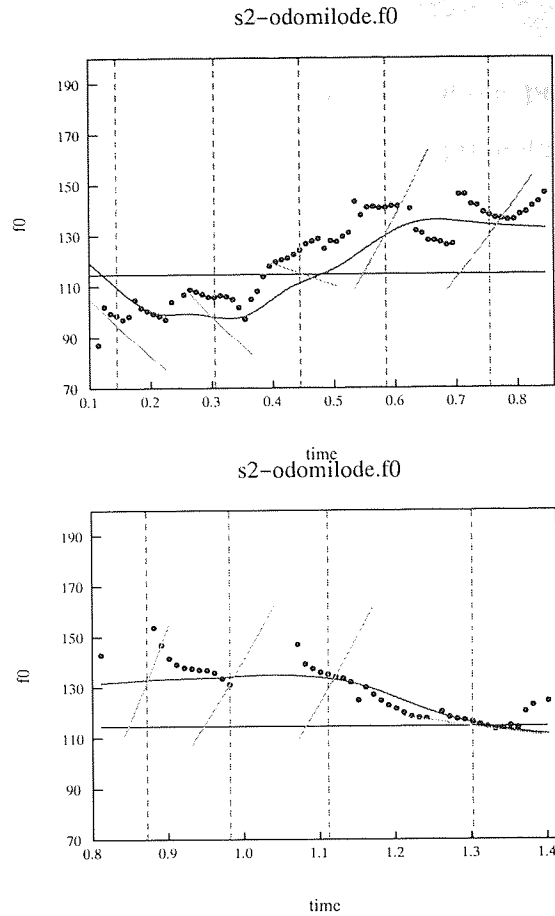


Figure 12.9: Stem-ML prediction of  $f_0$  and raw data for the two phrases of SY sentence “*Ôðòmi lóðé, Kó tó lə.*”. This data is in the test set; the model prediction is based on parameters derived from the training set, using syllable boundaries for this specific utterance.

physiologically motivated theories. Therefore, our R-Tree based prosody model can be expanded, and hence improved, when more phonetic and phonological knowledge become available in future. Furthermore, a functional R-Tree based prosody model can be implemented with a relatively small amount of data, as long as the phonological and phonetic description of the target language is adequate.

Detailed analysis of the model evaluations can be easily achieved in the R-Tree based prosody model. For example, the performance of the duration and intonation models can be determined independently before they are incorporated into the R-Tree. The ability to analyse the prosody dimensions independently makes it possible to determine the aspect of the model that needs improvements. Unlike our R-Tree based model, it is impossible to evaluate the performance of a Stem-ML model on each prosodic dimension.

## 12.6 Summary

In this chapter, we have discussed the implementation of our prosody model in terms of the duration and intonation dimensions. We first presented an overview of the model implementation strategy where the data structure and the alignment models are discussed. The waveform synthesis is then discussed for the implementation of the prosody model. How the model is used to predict the duration and  $f_0$  data for two sample sentences (a one-phase and a two-phase sentence) was also presented. We also discussed the implementation of a Stem-ML model based on our SY speech data.

In the next chapter, we shall discuss the evaluation of the two models with the view to identify the strengths and weaknesses of our model when compared with the Stem-ML model.

## Part V

# Model Evaluation, Summary and Conclusion

## Chapter 13

# Evaluation and discussion

In Chapter 12, we have presented the complete R-Tree and the Stem-ML prosody models. We have also illustrated how they are used to model the duration and  $f_0$  dimensions. An informal visual inspection of the intonation produced by the two models shows that they both follow the general pattern of the natural  $f_0$  contour and that the durations they predicted are equally within the range of the natural speech. However, the R-Tree based model places the tonal targets more accurately than the Stem-ML model. This is an expected result because the R-Tree model is more phonologically motivated than the Stem-ML model.

Another interesting observation on the performance of the two models is the division of the intonation contours into two prosodic phrases or units when the utterance contains two phonological phrases. A visual observation of the speech spectrogram and waveforms of the synthesised speech parameters also shows minor variations arising from the interaction of invariant consonant and vowel target parameters under different prosody conditions.

The purpose of this chapter is to evaluate these models. The aim of the evaluation experiment is to compare the performance of our proposed approach with a standard prosody modelling approach with the aim to determine its performance in the context of modern TTS approaches. The Stem-ML model has been used in intonation and prosody model for tone languages such as Mandarin and Cantonese Chinese (*Kochanski and Shih*, 2003; *Kochanski et al.*, 2003b).

The evaluation is divided into two major types: (i) quantitative, and (ii) qualitative.

In the quantitative evaluation, the aim is to investigate how accurately the models predict the data. In the case of the qualitative evaluation, the aim is to see how native speakers judge the overall quality of the synthesised speech produced by the two models. As stated earlier in this thesis, the intelligibility and perceived naturalness of synthetic speech strongly depend on the prosodic quality.

### 13.1 Experimental data preparation

Our experimental data contain thirty isolated sentences of varying complexities. Their selection was based on the following criteria: (i) tone combinations, (ii) syllable types and length, (iii) phrasal structure (i.e. one and two phrase sentences). With respect to tone combinations, for example, we selected some sentences in which all syllables carry the same tones, e.g. “Ó tún sáré wá” (meaning “[*He again run come*] *He ran here again.*”) while others contain syllables with various combination of tones, some beginning and ending with a particular tone sequence, e.g. HHH, LLL or MMM.

An attempt was made to achieve a balance between the phonetic types of syllables in a sentence, but this was not possible in most cases due to the sparse nature of SY syllable type distribution. The length of the sentences used in this experiment is limited to fifteen syllables in order to reduce the cognitive load of the stimuli.

Fifteen out of the thirty sentences come from the training set and the remaining fifteen are from the test set. The fifteen sentences in the test sets are further divided into two groups. The first test data group consists of ten sentences. The syllables that make up these sentences are in our syllable database which was used for developing the R-Tree based prosody model. The remaining five sentences, which form the second test group, are composed from a number of syllables that are not in our syllable database. The reason for composing the second test group is to see how well the implemented models are able to extrapolate from known to unknown data.

All test sentences are statement sentences. Three types of stimuli were prepared for each of the sentences: (i) the speech synthesised using the parameters computed by the R-Tree model, (ii) the speech synthesised using the parameters computed by the Stem-ML model, and (iii) the natural speech spoken by an adult male native speaker of SY.

The synthesised stimuli were created by replacing the *DurationTiers* and *PitchTiers* of the natural speech with the ones computed by each model and applying the PSOLA re-synthesis function in the *Praat* speech processing software (Boersma and Weenink, 2004).

For easy identification, the stimuli were labelled using codes:  $FLxyz$ ,  $SMxyz$ ,  $NAxyz$ , where  $xyz$  is the sample identification number. In the sample identification number,  $xy$  is the sample number and its value ranges from 1 to 30 while  $z$  is an identifier for the sample set. When  $z$  equals 1, it implies that the sample is in the training set; when  $z = 2$ , it implies that the sample is in the first test set, i.e. the group of ten test set. When  $z = 3$ , it implies that the sample is in the second test set, i.e. the group of five test set. The codes FL, SM and NA indicate the origin of the speech sample, i.e. generated by the R-Tree model, or the Stem-ML model, or it is natural. This process yields a total of 90 different stimuli.

## 13.2 Quantitative evaluation

Two parameters are used for our quantitative evaluation: the Root Mean Square Error (RMSE) and Pearson's Correlation (Corr) (Petruccelli et al., 1999). The RMSE measures how far apart two intonation contours are, while Pearson's Correlation measures how closely the synthetic  $f_0$  contour relates to the natural one. RMSE measures the distance between two contours on the time axis regardless of the  $f_0$  contour shape. Pearson's correlation, on the other hand, measures the degree to which variables are linearly related.

On the  $f_0$  dimension, the Stem-ML model fits the first and second test sets with an RMSE of  $17.00Hz$  and  $18.80Hz$  respectively. The fit to the training set has a RMSE of  $12.00Hz$ . The R-Tree model fits the first and second test sets with an RMSE of  $17.40Hz$  and  $17.30Hz$ ; the fit of the training set has an RMSE of  $16.50Hz$ . This result implies that the  $f_0$  contour predicted by the Stem-ML model is closer to the natural  $f_0$  contour when we consider the training set. However, the accuracy of the Stem-ML based model is better, although not significantly, when we compare the test data sets. This confirms the well known fact that qualitative approaches, i.e. the Stem-ML

Table 13.1:  $f_0$  contour evaluation

Model	Data set	RMSE(ms)	Corr.
R-Tree based	Training data set	16.50 (2.30)	0.77 (0.11)
R-Tree based	First data set	17.40 (2.11)	0.68 (0.06)
R-Tree based	Second data set	18.35 (1.72)	0.57 (0.05)
Stem ML-based	Training data set	12.00 (3.11)	0.85 (0.13)
Stem ML-based	First data set	16.56 (2.78)	0.76 (0.09)
Stem ML-based	Second data set	18.80 (1.89)	0.61 (0.07)

based model in this case, are strong at modelling the training data but are weak at extrapolating to test data.

The correlation for the Stem-ML model was 0.85 (0.13) for the training set. For the first and second test set it was 0.76 (0.09) and 0.61 (0.07) respectively. The correlation for the R-Tree model is 0.77 (0.11) for the training set. For the first and second test sets, it is 0.68 (0.06) and 0.57 (0.05) respectively. This result indicates that the Stem-ML model produces a synthetic  $f_0$  contour that is more closely related to the natural one.

The correlation results indicate that Stem-ML models the training and test data more accurately than the R-Tree model. However, the standard deviation for the R-Tree based prosody model is smaller than that of the Stem-ML model. This means that the R-Tree model produced a more consistent prediction than the Stem-ML model.

From this standpoint, the quantitative analysis does not produce a conclusive result as we might have hoped. The reason may be that the two quantitative analyses measure two different aspects of the  $f_0$  contour: the RMSE measures how far apart the natural and synthetic  $f_0$  contours are while the correlation measures how closely the synthetic  $f_0$  contour relates to the natural one. If a few of the synthesised  $f_0$  points are far away from the natural  $f_0$  contour, the RMSE may be quite high while the correlation, which is a more cumulative measure, may not be affected as much.

Much of the error in the Stem-ML based prosody model can probably be accounted for by segmental effects (*Silverman, 1987; Dusterhoff, 2000; Kochanski et al., 2003b*), due to changes in the vowel and the syllable onset consonant. Generally speaking, the

worst-fitting syllables are those with the largest and fastest  $f_0$  excursions. These are conditions where Stem-ML's approximations between templates and the realised pitch curve is furthest from the actual perceptual metric. However, segmental effects or unmodelled differences in the strength of syllables may also play a role. The results of the fits for Stem-ML models are generally similar to other Stem-ML based intonation models (Kochanski *et al.*, 2003b; Kochanski and Shih, 2003).

The incorporation of perceptual information into the R-Tree based model, by way of stylisation of the  $f_0$  curves, may have accounted for its consistent results when compared to that of Stem-ML. In addition, despite the apparent success of the Stem-ML model at capturing some aspects of SY intonation, it is difficult to relate these results to linguistic phenomena because the model does not take the phonological rules of SY into account.

Another interesting aspect of the results in this experiment is that the RMSE for the  $f_0$  dimension for the two models is less than the JND (Just Noticeable Difference) of 22Hz reported for SY by Harrison (2000). This implies that, although the errors are a bit large, they are unlikely to create a significant perceptual disparity from the natural speech since they are less than the minimum  $f_0$  value range that is noticeable by a native speaker.

However, these quantitative evaluations cannot determine whether variation in the synthetic contour makes the synthetic speech sound any more or less natural. They only measure how much the two contours vary. Therefore, there was a need to perform qualitative evaluation.

### 13.3 Qualitative evaluation

An important aspect of synthetic speech is how easy it is to understand (Hawkins *et al.*, 2000; Sun, 2002). The ease of understanding a spoken message is, however, influenced by the prosody of the speech. The qualitative evaluation aims to provide an insight into the opinion of the potential users of a TTS system concerning the quality of the synthesised speech. Substantial efforts have been made to develop good methodologies for evaluating the quality of synthetic speech (e.g. Huggins and Nickerson (1985);



*Monaghan and Ladd (1990); Benoît et al. (1996); Bradlow et al. (1996); Zera (2004); Viswanathan and Viswanathan (2005)).*

Our qualitative evaluation experiment has two aims: (i) determining how much of the synthetic speech is understandable (i.e. its intelligibility), and (ii) assessing how pleasant it is to listen to the synthesised speech as a whole (i.e. its naturalness). It has been shown that people's perception of synthetic speech also depend on para-linguistic factors such as the gender of the voice being mimicked as well as the context of application (*Mullennix et al., 2003*).

The primary approach has been to use a subjective method whereby native speakers of the language listen to synthetic speech and rate it on a scale of one to five. In the following subsection, we summarise the result of the intelligibility and naturalness evaluations of the two prosody models discussed in Chapter 12.

To prepare the data for qualitative evaluation, thirty sentences were selected from the ninety sentences discussed above (cf. Section 13.1). Ten of the sentences were from the training set, ten from the first test set and five from the second test set. The remaining five are natural speech. The participants were informed that the experiment dealt with the quality of synthetic speech. The sentences were presented to each of the participants in a random manner. Each sample is played to the participant as many times as they wanted. The participants were not informed about which of the speech samples are natural or synthesised.

Seventeen native speakers of SY volunteered to take part in the evaluation experiment, ranging in age between 23 and 45 years old. To ascertain their hearing ability, they were all subjected to an initial screening process. That process involves playing some natural speech sound to them and asking them to transcribe or repeat what they heard. Those who failed to produce 100% accuracy in this test were excluded from the evaluation experiment. As a result, fourteen of them were selected for the evaluation. All fourteen participants took part in the evaluation of the training and first test set evaluation, but only ten of them were also able to participate in the evaluation of the second test set. Each participant took about 45 minutes to evaluate the speech. In each case, the intelligibility evaluation was done first. After a five-minute break the naturalness evaluation followed.

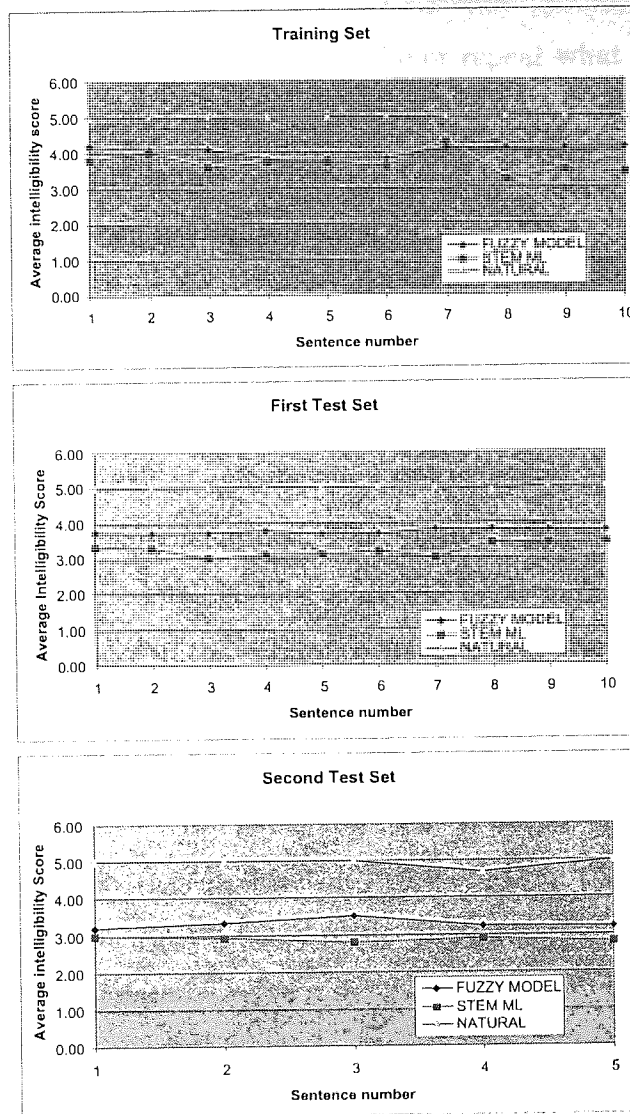


Figure 13.1: Result of intelligibility evaluation

### 13.3.1 Intelligibility evaluation

The aim of the intelligibility test is to ascertain whether native speakers can associate some meaning to synthesised speech after listening to the sound. Our intelligibility test is designed to establish how well the listeners identify the syllables in an utterances. This intelligibility evaluation is adopted because we feel that the syllable is the most important perceptual unit in an SY utterance. If the listeners are able to identify all syllables in a speech, then they can discern the underlying words, phrases and sentences.

The intelligibility test we adopted here is a transcription error test. During the test,

the listeners are expected to be able to transcribe or repeat what they have heard. In our intelligibility evaluation, we use a strategy in which the number of syllables that are correctly recognised is weighted against those that are wrongly recognised. The result of this weighting is then normalised on a scale of 5. The computation is achieved using the formula:

$$Intelligibility = \left( \frac{T_{All} - T_{Wrong}}{T_{All}} \right) \times 5.0 \quad (13.1)$$

where  $T_{All}$  is the total number of syllables in a sentence and  $T_{Wrong}$  is the number of syllables that had been wrongly identified. The formula ensures that if all the syllables are wrongly transcribed, the intelligibility score will be zero. The intelligibility scores will be five if all the syllables in the utterance are correctly transcribed.

On the training set, R-Tree based model was rated to have an intelligibility rating of 4.0, whereas the Stem-ML model was rated 3.6 (see Table 13.2). A sign test statistics (cf. Section 10.3.3) shows that the listeners preferred the synthetic speech generated using the R-Tree based model (significant at  $p \leq 0.05$ ) than that generated using the Stem-ML based model. The R-Tree based model also outperforms the Stem-ML model on the first and second test data by being rated 3.7 and 3.3 as opposed to 3.3 and 2.9 for the Stem-ML model. However, there is no statistically significant evidence ( $p > 0.05$ ) that the speech generated using the R-Tree based model is more preferred than that generated using the Stem-ML in this case. These results show that the synthetic speech produced using the R-Tree based prosody model is slightly more intelligible than that produced using the Stem-ML based prosody model. Noting that the average intelligibility score for the natural speech in this test is 5.0 and that for the test set is 4.9.

Table 13.2: Results for intelligibility evaluation

	Training data	first test data	Second test data
R-Tree based model	4.0 (0.15)	3.7 (0.05)	3.3 (0.12)
Stem-ML based model	3.6 (0.31)	3.2 (0.16)	2.9 (0.08)
Natural speech	5.0 (0.01)	4.9 (0.03)	4.9 (0.03)

*In this and the subsequent tables, the number in parenthesis is the standard deviation.*

Table 13.3: Qualitative evaluation scores

Value	Description
5	Perfect, indistinguishable from natural speech quality
4	Very good
3	Average
2	Poor
1	Weak or not acceptable

### 13.3.2 Naturalness evaluation

The average Mean Opinion Score (MOS) and the standard deviation for each model over the thirty stimuli were also computed. In this case, the participants were asked to rate their overall impression of the speech quality in terms of how close it is to human speech. They were asked to rate the quality of the stimuli on the 5-point MOS scale as described in Table 13.3. The participants were not informed which speech was natural or synthesised.

As shown in Table 13.4, in the training set, both the R-Tree based model and the Stem-ML model score the same MOS of 3.1, but the result of the R-Tree model has a lower standard deviation. On the test set, however, the results show that the R-Tree base model received the higher average rating. While the first and second test sets were rated as 2.5 and 2.2 respectively for the R-Tree based prosody model, they are rated 2.3 and 1.9 for the Stem-ML model.

The result of a sign test (cf. Section 10.3.3) statistics on the first test set shows that there is no evidence ( $p > 0.05$ ) that the R-Tree based model is preferred over the Stem-ML model. The same result is obtained for the second test set. In a further analysis, we found that there is no statistically significant evidence ( $p > 0.05$ ) that listeners preferred the synthetic speech generated by the Stem-ML model over that of the R-Tree based model. This shows that the R-Tree based model is at least not significantly worse than the Stem-ML model in terms of naturalness.

When the MOSs of each participant are examined individually, all but three of them rated the Stem-ML model higher than the R-Tree based model. Some participants

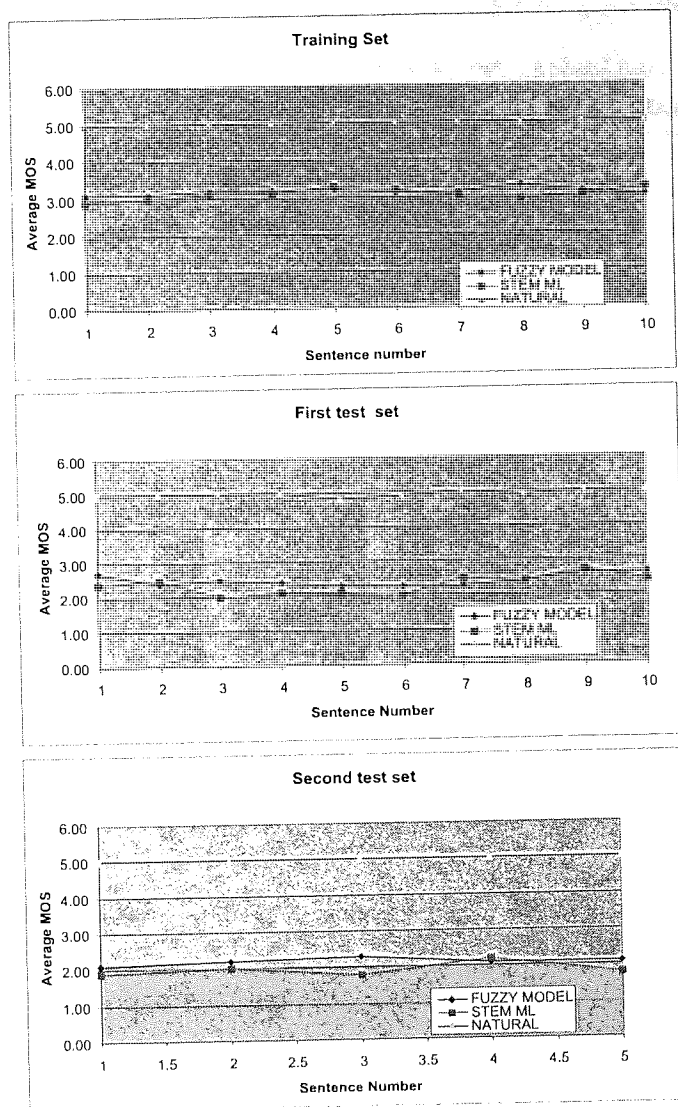


Figure 13.2: Result of naturalness evaluation

tend to be strict while others tend to be more lenient with judging the quality of the speech, particularly those that have discontinuous points. This can be seen from the overall standard deviation for each of the prosody models over all participants. The calculated mean of the SD shows that the FDT model is more consistent than the Stem-ML model since it has a smaller mean SD. A more consistent system should yield a smaller standard deviation.

Table 13.4: Results for naturalness evaluation

	Training data	first test data	Second test data
R-Tree based model	3.1(0.08)	2.5(0.02)	2.2(0.08)
Stem-ML based model	3.1(0.13)	2.3(0.03)	1.9(0.17)
Natural speech	5.0 (0.00)	4.8(0.01)	4.7(0.02)

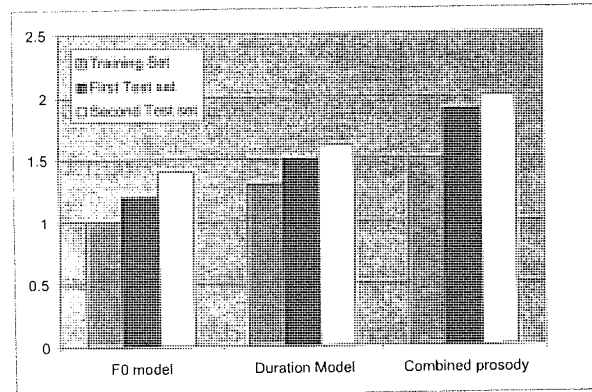


Figure 13.3: G value evaluation

## 13.4 Discussion

The approach adopted in this research suggests that each dimension of prosody, i.e. duration and  $f_0$ , must be viewed as a single entity at the phonological level. However, as we move from phonological towards the phonetic and acoustic domain, it is more appropriate to model each dimension individually. One motivation for this is that each of the prosody dimension has its own peculiar characteristics which necessitate the application of a different modelling approach. The intonation dimension, for example, can be modelled using continuous functions over the voiced portion of a syllable. Duration dimension, on the other hand, can be treated as comprising of discrete entities for describing the span of syllables, words, phrases or even sentences.

To ascertain the synthetic speech quality, let  $X$  be the MOS for the natural utterance and  $Y$  be the MOS for the synthesised utterance. Then the difference  $G = Y - X$  would be a good indication of the relative improvement or degradation as a result of the prosody parameter computed by the model. Three groups of data were selected for this analysis: (i) the training set, (ii) the first test set and (iii) the second test set. In Figure 13.3, we plotted the mean value of  $G$  for synthetic speech generated using the

duration and  $f_0$  model as well as those obtained by the combination of the two.

It can be seen that the degradation is marginal but consistent. It is also observed that the duration dimension introduces the most degradation. With this result, it is clear that we need to further improve the model in general. The duration model deserves more improvement, since it has a significant impact on the overall quality of the synthetic speech and it incurs the higher error.

It is possible that some alignment shift may have occurred in the Stem-ML model. This is because, with the Stem-ML model, it is difficult to determine how the  $f_0$  contour aligns with the syllable. Unlike our R-Tree based model, which generates the  $f_0$  contour on a syllable by syllable basis, the Stem-ML model generates a smooth  $f_0$  contour over an entire phrase. This makes it difficult to determine syllable and word boundaries.

It may be argued that the number of participants in the qualitative evaluation is small. However, as shown by the results of the evaluations, the overall rating of the synthetic speech is consistent. This suggests that if a larger population is used we should not expect much changes in the overall evaluation ratings. It may also be argued, based on the Just Noticeable Difference, that the differences in the intonation pattern predicted by the models were too subtle to produce a significant difference in perception. The perceptual evaluation, however, suggests that this is not the case and it also indicates that the important issue is not the pattern of intonation but whether the pattern of intonation incorporates the systematic phonological and phonetic structure of the speech sound.

We observed that the error incurred in modelling each of the prosody dimensions is lower than that incurred when the dimensions are combined into an R-Tree. This is expected because the individual error in the part should sum up to the error in the whole. However, the error incurred in the combined model is not an additive or a multiplicative error. By this, we mean that the error in all of the dimensions is not a sum or a product of the error in the individual dimension. In fact, the error incurred in the combined model is about 15% worse than that in the dimension with the highest error.

For example, the RMSE for the training set in the duration model is  $14.59ms$  (cf. Chapter 11), while that in the  $f_0$  contour model is  $15.56Hz$  (cf. Chapter 10) but

the RMSE in the combined prosody model is  $16.50Hz$  for the  $f_0$  dimension and  $17.23ms$  for the duration dimension. This quality of the model is particularly advantageous since we can be sure that the aggregate error incurred by combining the dimensions will not be astronomically high.

We observe that, in both models, most of the intelligibility errors arise from syllables in which the onset and rhyme are from the same or similar phonetic classes. For example, the syllable */wa/* and */ya/* are usually confused with each other because the onset consonant */w/* and */y/* are semi-vowels. It is noted that, listeners seem to use the semantic structure, rather than the acoustic signals, of the sentence to determine the correct syllable in some cases. Also, CV and CVn syllables in which the nuclei have the same vowel are easily confused by the listeners. For example, in most cases, *ba* is perceived as *ban* and *bə* is perceived as *bən*.

This kind of errors accounts for the lower intelligibility score for the second test set discussed in the above evaluation. We have not been able to determine how best to model this aspect at the moment, but an approach will be to introduce a separate fuzzy model for these exceptional cases. The fuzzy rule will then be invoked when this type of syllables is to be synthesised. We anticipate that integrating this rule of exception with the other already implemented rules will pose some other challenges.

## 13.5 Summary

In this chapter, we have presented the evaluation of our proposed prosody model and discussed its performance. Stem-ML is a contemporary prosody modelling technique that has been shown to be very successful in modelling prosody for tone languages. We have shown that our R-Tree based prosody modelling approach models the data less accurately than the Stem-ML model. However, our model performed slightly better than the Stem-ML, at extrapolation from the training data to the test set.

In addition, our R-Tree based modelling approach produced a better speech quality than data-driven methods. In essence, we can say that our approach preserves much of the natural prosodic structure of speech. This has been attributed to the incorporation of phonological knowledge into our prosody model which may have introduced syste-



## CHAPTER 13. EVALUATION AND DISCUSSION

matic details into the synthetic speech, hence increasing its robustness, and in turn, its perceptual quality.

Further analysis suggests that improvements to our model are required and that the duration dimension introduced the most degradation to the quality of the synthetic speech. This analysis was made possible by the modular holistic approach to prosody modelling using R-Tree. Hence, in this chapter, we have also confirmed the validity of our hypothesis in adopting a modular holistic approach.

## Chapter 14

# Summary, conclusion and further work

This thesis reports the foundations of a work whose long term aim is to use computational models for implementing an accurate prosody system for tone language text-to-speech synthesis. The Standard Yorùbá (SY) language is the focus of the present research. The main philosophy underlying the development of the prosody model is:

- To model speech prosody, there is the need to integrate and use both speech data and expert knowledge.
- A holistic view of modelling various prosody dimensions, yet allowing individual dimensions to be implemented separately and then integrated to form the required synthesised speech output, is a more appropriate approach to prosody modelling.
- Fuzzy logic is a potent tool for modelling speech sound parameters.

We have confirmed that good quantitative performance, in terms of root mean square error (RMSE), means square error (MSE) and correlation, does not always translate to good synthetic speech quality. We have also confirmed that good quality speech, in terms of intelligibility and naturalness, requires that the perceptually significant portion of a speech signal be accounted for in a prosody model. These findings can be explained in the background of the well known “redundancy” of the speech signal, whereby a phone can be signalled by a number of co-occurring acoustic proper-

ties. Hence, accuracy based on speech data alone will not reflect the multidimensional properties of the linguistic structure from which the speech signal is constructed.

It has also been suggested that listeners have a threshold, which can be related to the Just Noticeable Difference (JND) (*Harrison, 2000*), within which the perceptual quality of speech is tolerant to error in the speech signal. If this threshold is not exceeded, the synthesised speech signal may still retain adequate perceptual quality.

## 14.1 Summary

This thesis has presented a prosody model for text-to-speech synthesis (TTS). The model assumes that the abstract and realised forms of intonation and other dimensions of prosody should be modelled within a modular and holistic view. That idea is conceptualised within a unified framework which is implemented using the Relational Tree (R-Tree) technique. The R-Tree is a sophisticated data structure for representing a multi-dimensional waveform in the form of a tree.

The R-Tree for an utterance is generated in two steps. First, we generate the abstract structure of the waveform, known as the Skeletal Tree (S-Tree). The algorithm for generating the S-Tree is derived from the tone phonological rules for the target language. Second, we compute the numerical values of the perceptually significant peaks and valleys on the S-Tree using fuzzy logic based models. The resulting points are then joined by applying interpolation techniques. The actual intonation contour is synthesised by Pitch Synchronous Overlap Technique (PSOLA) using the *Praat* software.

The realisation of the intonation dimension is achieved by proposing a mathematical model for generating the  $f_0$  contours. The fuzzy logic model structure and parameter identification were obtained. Representing an  $f_0$  contour using the peaks and valleys of the  $f_0$  curve of the component tones in an utterance and using a fuzzy logic based model to compute their numerical values provide an  $f_0$  description which is essential for capturing both the linguistic and acoustic aspects of intonation. The reason for this is that it enables the fuzzy model to generate  $f_0$  contours from a linguistically meaningful description which is directly related to the phonological system of the language. As a

result, our model is easy to interpret, implement and adapt.

Judging from the number of parameters and variables involved in our model, one may argue that the model is complex. The simplicity of the model structure however, outweighs any complexity stemming from the number of variables used in the model. That is because the variables and parameters are linguistically motivated, and contribute to the explicit representation of the prosody phenomena, therefore making the model easier to interpret.

In addition to extending the expressiveness of the R-Tree, the incorporation of a fuzzy model into a symbolic method (i.e. the R-Tree approach) increases the flexibility and robustness of our model. In particular, it allows a set of linguistic rules to be related by the dual concept of fuzzy implication and compositional rule of inference.

To realise the duration dimension, we applied the fuzzy decision tree technique. We found that, when compared with the CART-based approach, our FDT-based duration model better captures those salient aspects of the speech signal that have greater perceptual significance. In this regard, the FDT model is more appropriate for modelling duration in the context of TTS applications for tone languages. Furthermore, fuzzification and global optimisation provide a continuous representation with the flexibility necessary to reproduce duration patterns at a finer granularity.

We have demonstrated the applicability of our model using the Standard Yorùbá (SY) language. Our preliminary evaluation results have confirmed the effectiveness of our holistic approach to prosody modelling. Our prosody model is able to generate a set of  $f_0$  patterns that are easy to align with syllables and that are within the “perceptual tolerance” of native speakers. This is particularly important for TTS applications, since the aim in such applications is to produce synthetic speech with acceptable perceptual quality and not to model the speech data with the highest numerical accuracy.

Furthermore, our approach provides a linguistically motivated computational mechanism that can form the basis of a system for asking appropriate questions about speech prosody. Our model can help to advance the knowledge in the field of speech science as it can serve as an instrument for clarifying acoustic patterns into distinct linguistic classes. It will also facilitate the analysis of speech representation as well as the generation of rules and algorithms which will go some way to realising a fully

explicit and computationally testable model of linguistic theory (*Gibbon, 2001*).

Given that a prosody model simulates the observed data well, there are three questions that are much debated in the literature. The first is whether or not the model's primitives, i.e. the representation of prosody elements, are appropriately chosen to reflect the perceptually significant aspects of speech signal (*Botinis et al., 2001*). The second issue is whether the prosody model is linguistically or phonetically plausible (*Thorsen, 1985; Taylor, 1994*). The third issue, which is an offshoot of the second, is whether or not the model can be interpreted physiologically (*Cohen, 1995; Kochanski and Shih, 2003*). Our model contributes to the first two issues in this debate.

The results obtained from our model show that the tonal primitives derived through stylisation and standardisation techniques can adequately represent perceptual cues in an  $f_0$  curve. The results also show that in order to model the perceptually significant aspects of speech prosody, there is the need to establish a link between the linguistic and acoustic aspects of speech. Our results also confirm the findings in other work, e.g. *Möbius (2003)*, which has shown that data-driven approaches cannot sufficiently model the qualitative aspects of speech prosody, which are difficult to represent in a speech corpus due to sparse data.

Our results are in line with the finding of other researchers (e.g. *Ladd (1990); Clark and Dusterhoff (1999); Campbell (2000)*) that an accurate prediction of speech data is necessary, but not sufficient, for the perception of speech prosody. This indicates that a model must not only simulate a given data set accurately, it must also be able to represent and accurately predict aspects of the prosody phenomena (e.g. final lowering, L lowering, etc.) that are difficult to model from data due to their sparse nature. In addition to this, a number of these phenomena, e.g. downtrend, are composed of a hierarchy of events that are not available to data-driven models and cannot be easily captured automatically by such models. This necessitates the need to integrate speech data and expert knowledge in prosody modelling.

The idea here is not to reject the data-driven approach but rather to point out that, in our opinion, rules based on phonetic and phonological knowledge should first be used to establish a systematic structure for the speech sound. Such a structure will then be used to guide the data collection process as well as to focus the scope of the

speech corpus. A data-driven approach can then be designed based on the collected speech corpus. The result of evaluating such a data-driven model can then act as a feedback for collecting better data and further improving the system.

We therefore propose a prosody modelling paradigm whereby the rule-driven and data-driven approaches are integrated. In this modelling paradigm, the rule-driven approach will serve as a mechanism for identifying the relevant features to be modelled, while the data-driven will serve as a mechanism for learning such features automatically from the data. The process will be iterative, in that the rule-driven and data-driven processes will be repeated alternatively until a desirable prosody model is obtained.

We can summarise, therefore, that our R-Tree based prosody model exhibits high comprehensibility, and that the incorporation of the fuzzy set and approximate reasoning methods provides a natural means to deal with continuous domains and subjective linguistic descriptions which are the characteristics of prosody data and phenomena.

## 14.2 Contribution to knowledge

The study presented in this thesis has made a number of contributions to knowledge in the areas of computer text-to-speech synthesis and intelligent systems engineering. The specific contributions are as follows:

**R-Tree based prosody model** We have proposed, designed and implemented a Relational Tree based prosody model. The R-Tree approach is popular in intelligent systems engineering (*Ehrich and Foith, 1976; Shaw and Defigueiredo, 1990*). Our model is unique in the sense that this work is the first to apply the R-Tree approach to prosody modelling.

**Integration of fuzzy logic model into R-Tree** The R-Trees used in intelligent systems engineering are commonly applied in analysing waveforms. Analytical techniques are used to compute points in the R-Tree. In this work, we employ fuzzy logic techniques for Relational Tree realisation.

**Fuzzy control rule in modelling intonation** Data-driven and rule based approaches (*Sakurai et al., 2003*) are widely used in intonation modelling in the literature.

Based on our idea of modelling the linguistic and numerical aspects of some intonation phenomena within a unified framework, we introduce a fuzzy logic based approach. Fuzzy logic is very popular in intelligent process control, but this work is the first to apply fuzzy logic to intonation modelling.

**Fuzzy decision tree in duration modelling** Classification and Regression Tree (*Riley, 1992; Chung and Huckvale, 2001; Donovan, 2003*) is a popular tree based duration modelling approach reported in the literature. Our use of a fuzzy decision tree, in this context, is pioneering.

**First work on Yorùbá prosody model** The work presented in this thesis represents a pioneering work in the development of a prosody model for a standard Yorùbá (SY) TTS system. Our prosody model has the potential to be applied to tone language speech recognition as well as to the study of Standard Yorùbá prosody in general.

In order to achieve the contribution stated above, the following tasks were accomplished:

1. A relational tree model was proposed within which the abstract and realised aspects of prosody can be implemented. To implement the model, two tasks were identified: (i) the generation of a skeletal tree (S-Tree) which describes the abstract waveform, (ii) the computation of the dimensions of the target waveform using the generated S-Tree as a guide.
2. An algorithmic implementation of S-Tree generation, based on the tone phonological rules of the SY language has been developed.
3. A small SY language resource was created for the prosody modelling. An SY speech database was also created and annotated.
4. A computation of the fundamental frequency dimension using fuzzy rules has been implemented and tested. The model structure was determined based on linguistic and phonetic experts' knowledge documented in the literature. The parameters of the model were determined using data collected with respect to the intonation of a native speaker.

5. The duration dimension was implemented using the fuzzy decision tree method. This model was compared with an equivalent CART-based model using the same training data. The fuzzy logic approach was found to produce better speech quality although it does not model the data as accurately as the data driven approach.
6. The results of the fundamental frequency and the duration modelling were incorporated into the S-Tree and the resulting prosody model was evaluated. A comparison between the implemented model and a Stem-ML model suggests that our R-Tree based approach produces synthesised speech that is more natural.

### 14.3 Extension of the model for other tone languages

In our opinion, it is relatively straight-forward to apply our model to the generation of prosody for other tone languages. In order to achieve that, our approach for modelling the fundamental frequency and the duration dimensions needs to be adjusted. The following adjustments are required for the fundamental frequency dimension modelling:

1. Determine the most perceptually acceptable approximation of the  $f_0$  curve for each tone in the target language using the stylisation and standardisation techniques (e.g. *tHart* (1991); *d'Alessandro and Mertens* (1995)) (cf. Chapter 9).
2. Identify and label the peak(s) and valley(s) on the stylised  $f_0$  curves (cf. Chapter 9).
3. Derive an algorithm for determining the highest peaks and lowest valleys in a sequence of tones based on the tone phonology of the target language as discussed in Chapter 9.
4. Design the fuzzy model by determining the model structure and parameters as presented in Chapter 10.
5. Evaluate and iteratively improve the fundamental frequency dimension.



The following adjustments are required for the duration dimension modelling:

1. Determine the factors affecting duration in the target language. A syllable based approach is more desirable in this task.
2. Design a fuzzy decision tree based on the factors affecting duration and obtain the model's parameters on the data collected for the target language as discussed in Chapter 11.
3. Implement and evaluate the duration model and improve on it iteratively until the desired result is achieved.

Other issues that relate to implementing the R-Tree based prosody model for a different language is that of tonal alignment. It may be argued that the issue of tonal alignment is yet to be adequately addressed in the context of our model. That task requires a more systematic and comprehensive approach and suggests the need to tap into the phonetic and phonological knowledge of the SY language. The task is a complex one which requires an interdisciplinary action of research. We hope to collaborate with phoneticians and phonologists on the SY language in the near future in order to address the problem and further improve our model.

## 14.4 Future work

At present, we have tested our model on short sentences containing not more than two intonation phrases, each with no more than fifteen syllables. Our observation of SY text materials, which includes newspapers and textbooks, suggests that about 85% of the sentences fall into this category. Therefore, the statement sentences used in the design and evaluation of our model have a simple structure (i.e. sentences of up to two phrases). Despite this, we have established a framework within which the model can be systematically analysed, expanded and further improved upon. In addition, we hope to expand the capacity of our model to accommodate longer and more complex sentence structures.

In the near future, we also hope to use our model to predict and test hypotheses about intonation phenomena. For example, *Connell and Ladd* (1990) have suggested that sentence mode does not have significant impact on the realised  $f_0$  contour of SY utterances. *Láníran and Clements* (2003) also suggested that Yorùbá speakers use

different strategies to generate intonation contours. These hypotheses have not been tested in the TTS context. It is desirable that these and other intonation hypotheses are tested using our model. The result of such an investigation will help us to fine-tune our model as well as to further contribute to the knowledge on modelling tone-language prosody.

Our goal has been to generate a complete model for synthesising speech prosody for Standard Yorùbá in particular and tone languages in general. In this respect, we will need to incorporate an intensity dimension into the R-Tree. Among other things, we will also need to address the following issues in a more comprehensive manner:

1. The development of a better method for integrating all the prosody dimensions into our R-Tree model.
2. The need to develop a more comprehensive SY speech database and language resource.
3. The need for further study on the alignment problem and the need to come up with a model which is capable of solving that problem.

A well prepared speech corpus is very important in the development of a speech synthesis system. The use of a hand-labelled corpus restricted the size of our speech database but it does not significantly affect the scope of the data we required to achieve the preliminary results that are presented here. It will be desirable to label our speech corpus automatically. If that can be done accurately, it will reduce the cost and time of developing a large speech corpus for prosody modelling. However, such a fully automatic system is not yet available (*Ida, 2002*). If any of the currently available semi-automatic tools are used, the problem still remains that we will need to go through the entire speech corpus manually to validate the data. Moreover, the development of a fully automatic speech corpus annotation system is still an active research problem to which we hope to contribute in the near future.

# Bibliography

- Adéwólé, L. O., The Yorùbá high tone syllable revisited, in *Work in progress*, 19, pp. 81–90, Department of Linguistics, University of Edinburgh, Edinburgh, 1986.
- Adéwólé, L. O., The categorical status and the function of the Yorùbá auxiliary verb with some structural analysis in GPSG, Ph.D. thesis, University of Edinburgh, Edinburgh, 1988.
- Aguero, P., and A. Bonafonte, Intonation modelling for TTS using joint extraction and prediction approach, <http://www.ssw5.org/papers/1045.pdf>, 2004, visited: Jul 2004.
- Aker, G., and M. Lennig, Intonation in text-to-speech synthesis: Evaluation of algorithms, *Journal of the Acoustical Society of America*, 77, 2157–2165, 1985.
- Akinlabí, A., Underspecification and phonology of Yorùbá /r/, *Linguistic Inquiry*, 24, 139–160, 1993.
- Akinlabí, A., and M. Liberman, Tonal complexes and tonal alignment, <http://www ldc.upenn.edu/myl/newNELS.pdf>, 2000, visited: Jun 2005.
- Alamolhoda, S. M., *Phonostatistics and Phonotactics of Syllable in Modern Persian*, 1<sup>st</sup> ed., The Finish Oriental Society, Helsinki, 2000.
- Allen, J., Overview of text-to-speech systems., in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sondí, pp. 741–790, Marcel Dekker Inc., New York, 1992.
- Allen, J., M. S. Hunnicutt, and D. Klatt, *From Text to Speech: the MITalk System*, Cambridge Press, Cambridge, 1987.
- Allen, J. B., How do humans process and recognize speech?, *IEEE Trans. on Speech & Audio Processing*, 2, 567–577, 1994.
- Anderson, D. R., D. J. Sweeney, and T. A. Williams, *Statistics for bussiness and economics*, 8<sup>th</sup> ed., South-western, United Kingdom, 2002.
- Arroyo-Figueroa, G., L. E. Sucar, and A. Villavicencio, Fuzzy intelligent system for the operation of fossil power plant, *Engineering Application of Artificial Intelligence*, 13, 431–439, 2000.

## BIBLIOGRAPHY

- Bákàrè, C. A., Discrimination and identification of Yorùbá tones: perception experiments and acoustic analysis, in *Language in Nigeria: essays in honour of Ayò Bámgbóṣé*, edited by K. Owólabí, pp. 32–67, Group Publishers, Ìbàdàn, 1995.
- Bámgbóṣé, A., *Yorùbá Orthography*, Cambridge University Press, Cambridge, 1965.
- Bámgbóṣé, A., The assimilation of low tone in Yorùbá, *Lingua*, 16, 1–13, 1966.
- Bámgbóṣé, A., *Fonólójì àti Gírámà Yorùbá*, University Press PLC, Ìbàdàn, 1990.
- Barbosa, P. A., and G. Bailly, Characterisation of rhythmic patterns for text-to-speech synthesis, *Speech Communication*, 15, 127–137, 1994.
- Bartkova, K., and C. Sorin, A model of segmental duration for speech synthesis in French, *Speech Communication*, 6, 245–260, 1987.
- Batůšek, R., A duration model for Czech text-to-speech synthesis, an International Conference on Speech Prosody 2002: <http://www.ipl.univ-aix.fr/sp2002/pdf/bastusek.pdf>, 2002, visited: Sep 2004.
- Beckman, M. E., and S. Jun, K-ToBI Korean ToBI labeling conventions, <http://www.humnet.ucla.edu/humnet/linguistics/people/jun/ktobi/k-tobi-V2.html>, 1996, visited: Jun 2004.
- Bellegarda, J. R., K. E. A. Silverman, K. Lenzo, and V. Anderson, Statistical prosody modelling: from corpus design to parameter estimation, *IEEE Trans. on Speech & Audio Processing*, 9, 52–66, 2001.
- Benoît, C., M. Grice, and V. Hazan, The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, *Speech Communication*, 18, 381–392, 1996.
- Beskow, J., Animation of talking agents., in *Proc. of Intl. Conf. on Auditory-Visual Speech Processing (AVSP97)*, pp. 149–152, Rhodos, Greece, 1997.
- Beskow, J., Trainable articulatory control models for visual speech synthesis, *International Journal of Speech Technology*, 7, 335–349, 2004.
- Bezdek, J. C., Fuzzy models: what are they, and why?, *IEEE Trans. on Fuzzy Systems*, 1, 1–6, 1993.
- Bird, S., Automated tone transcription, <http://www.idc.upenn.edu/sb/home/papers/9410022/941002.pdf>, 1994, visited: Apr 2004.
- Bird, S., Strategies for representing tone in African writing systems: a critical review, URL: <http://cogprints.org/2174/00/wll2.pdf>, 1998, visited: Jan 2003.
- Birkholz, P., and D. Jackèl, A three-dimensional model of the vocal tract for speech synthesis., in *Proc. of 15<sup>th</sup> Inter. Congr. of Phon. Sci.*, pp. 2597–2600, Barcelona, 2003.

## BIBLIOGRAPHY

- Black, A., R. Clark, S. King, Z. Heiga, P. Taylor, and R. Caley, The Festival speech synthesis system: system documentation, version 1.4.0, [http://www.cstr.ed.ac.uk/projects/festival/manual/festival\\_25.html#SEC112](http://www.cstr.ed.ac.uk/projects/festival/manual/festival_25.html#SEC112), 1999, visited: Apr 2004.
- Black, A. W., and P. Taylor, Automatically clustering similar units in speech synthesis, in *EuroSpeech '97*, vol. 2, pp. 601–604, 1997.
- Blackburn, C. S., and S. Young, Enhanced speech recognition using an articulatory production model trained on X-ray data, *Computer Speech & Language*, 15, 195–215, 2001.
- Boersma, P., and D. Weenink, *Praat*, doing phonetics by computer, <http://www.fon.hum.uva.nl/praat/>, 2004, visited: Mar 2004.
- Botinis, A., B. Granström, and B. Möbius, Development and paradigms in intonation research, *Speech Communication*, 33, 263–296, 2001.
- Bouzon, C., and D. Hirst, The influence of prosodic factors on the duration of words in British English, International Conference on Speech Prosody 2002: [www.lpl.univ-aix.fr/sp2002/pdf/fagyal.pdf](http://www.lpl.univ-aix.fr/sp2002/pdf/fagyal.pdf), 2002, visited: Aug 2004.
- Boyer, X., and L. Wehenkel, Automatic induction of fuzzy decision trees and its application to power systems' security assessment, *Fuzzy Sets & Systems*, 102, 3–19, 1999.
- Bradlow, A. R., G. M. Torretta, and D. B. Pisoni, Intelligibility of normal speech I: global and fine-grained acoustic-phonetic talkers characteristics, *Speech Communication*, 20, 255–272, 1996.
- Braunschweiler, N., Automatic detection of prosodic cues, Ph.D. thesis, Universität Konstanz, Germany, 2003.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Tree*, Wadworth, CA, 1984.
- Breining, C., A robust fuzzy-based step-gain control for adaptive filters in acoustic echo cancelation, *IEEE Trans. on Speech & Audio Processing*, 9, 162–167, 2001.
- Brinckmann, C., and J. Trouvain, The role of duration models and symbolic representation for timing in synthetic speech, *International Journal of Speech Technology*, 6, 21–31, 2003.
- Burnett, D. C., M. R. Walker, and A. Hunt, Speech Synthesis Markup Language Version 1.0 W3C Working Draft 02, <http://www.w3.org/TR/speech-synthesis/#S1.1>, 2002, visited: Jun 2004.
- Cahn, J. E., A computational memory and processing model for prosody, Ph.D. thesis, Media Arts and Science, School of Architecture and Planning, MIT, Cambridge, MA, 1998.

## BIBLIOGRAPHY

- Campbell, N., Autolabelling Japanese ToBI, <http://www.asel.udel.edu/icslp/cdrom/vol4/897/a897.pdf>, 1997, visited: May 2004.
- Campbell, N., Timing in speech: a multilevel process, in *Prosody: Theory and Experiment*, pp. 281–334, Kluwer Academic, Dordrecht, 2000.
- Campbell, N., and J. J. Venditti, J-ToBI: an intonation labelling system for Japanese, in *Proceedings of the Autumn meeting of the Acoustical Society of Japan*, vol. 1, pp. 317–318, Utsunomiya, 1995.
- Campione, E., E. Flachaire, D. Hirst, and J. Véronis, Stylization and symbolic coding of  $f_0$ : a quantitative model, <http://www.up.univ-mrs.fr/~veronis/pdf/1997escacampione.pdf>, 1997, visited: May 2004.
- Campione, E., D. Hirst, and J. Véronis, Automatic stylisation and symbolic coding of  $f_0$ : implementations of the INTSIT model, <http://www.up.univ-mrs.fr/~veronis/pdf/2000campione.pdf>, 2000, visited: May 2005.
- Carvalho, D. R., and A. A. Freitas, A genetic-algorithm for discovering small-disjunct rules in data mining, *Applied Soft Computing*, 2, 75–88, 2002.
- Castellano, P., and S. Sridharan, A two stage fuzzy decision classifier for speaker identification, *Speech Communication*, 18, 139–149, 1996.
- Castro, J. L., and M. Delgado, Fuzzy systems with defuzzification are universal approximator, *IEEE Trans. on Fuzzy Systems*, 26, 149–152, 1996.
- Chan, N.-C., and C. Chan, Prosodic rules for connected Mandarin synthesis, *Journal of Information Science & Engineering*, 8, 261–281, 1992.
- Chapentier, M. J., and M. G. Stella, Diphone synthesis using an overlap-add technique for speech waveforms concatenation, in *Intl. Conf. Acoust. Speech Signal Processing*, vol. 86, pp. 2015–2018, 1986.
- Chen, J. E., and K. N. Otto, Constructing membership functions using interpolation and measurement theory, *Fuzzy Sets & Systems*, 73, 313–327, 1995.
- Chen, S.-H., S.-H. Hwang, and Y.-R. Wang, An RNN-based prosodic information synthesiser for Mandarin text-to-speech, *IEEE Trans. on Speech & Audio Processing*, 6, 226–239, 1998.
- Chen, S.-H., W. H. Lai, and Y.-R. Wang, A new duration modelling approach for Mandarin speech, *IEEE Trans. on Speech & Audio Processing*, 11, 308–320, 2003.
- Cheng, Y.-C., and S.-Y. Lu, Waveform correlation by tree matching, *IEEE Trans. on Pattern Analysis & Machine Intelligence*, PAMI-7, 299–305, 1985.
- Chiang, T. H., J. S. Chang, M. Y. Lin, and K. Y. Su, Statistical word segmentation, *Journal of Chinese Linguistics*, 9, 147–174, 1996.
- Chou, F.-C., C.-Y. Tseng, and L.-S. Lee, A set of corpus-based text-to-speech synthesis technologies for Mandarin Chinese, *IEEE Trans. on Speech & Audio Processing*, 10, 481–494, 2002.

## BIBLIOGRAPHY

- Chung, G., Hierarchical duration modelling for a speech recognition system, Master's thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1997.
- Chung, G. Y., and S. Seneff, A hierarchical duration model for speech recognition based on the ANGEL framework, *Speech Communication*, 27, 113–134, 1999.
- Chung, H., Duration models and the perceptual evaluation of spoken Korean, in *International Conference on Speech Prosody*, pp. 219–222, Aix-en-Provence, France, 2002.
- Chung, H., and M. Huckvale, Linguistic factors affecting timing in Korean with application to speech synthesis, in *Proc. of EuroSpeech '01*, pp. 815–818, Aalborg, Denmark, 2001.
- Clark, R. A. J., Generating synthetic pitch contours using prosodic structure, Ph.D. thesis, University of Edinburgh, Edinburgh, 2003.
- Clark, R. A. J., and K. E. Dusterhoff, Objective methods for evaluating synthetic intonation, in *Proc. of the 6<sup>th</sup> European Conference on Speech Communication Technology*, vol. 4, pp. 1623–1626, Budapest, 1999.
- Clements, G. N., and S. J. Keyser, *CV Phonology: A Generative Theory of the Syllable*, Cambridge Press MIT, Cambridge, M.A., 1983.
- Cohen, A., Developing a nonsymbolic phonetic notation for speech synthesis, *Computational Linguistics*, 21, 568–575, 1995.
- Cohen, M. M., and D. W. Massaro, Modelling co-articulation in synthetic visual speech, in *Models and Techniques in Computer Animation*, edited by N. Magnenat-Thalmann and D. Thalmann, pp. 139–156, Springer-Verlag, Tokyo, 1993.
- Coker, C., and O. Fujimura, Model for the specification of the vocal tract area function, *Journal of the Acoustical Society of America*, 40, 1271, 1966.
- Coleman, J. S., “Synthesis-by-rule” without segments or rewrite-rules, in *Talking Machines: Theories, Models and Designs*, edited by G. Bailly, C. Benoit, and T. R. Sawallis, pp. 43–60, Elsevier, Amsterdam, 1992.
- Coleman, J. S., Polysyllabic words in the YorkTalk synthesis system, in *Phonological structure and forms: Papers in Laboratory Phonology III*, edited by P. A. Keating, pp. 293–324, Cambridge University Press, Cambridge, 1994.
- Collier, R., On the perceptual analysis of intonation, *Speech Communication*, 9, 443–451, 1990.
- Connell, B., Tone, utterance length and  $f_0$  scaling, in *International Symposium on Tonal Aspects of Languages*, Beijing, 2004, visited: Jun 2005.
- Connell, B., and D. R. Ladd, Aspect of pitch realisation in Yorùbá, *Phonology*, 7, 1–29, 1990.

## BIBLIOGRAPHY

- Connell, B. A., J. T. Hogan, and A. J. Rozsypal, Experimental evidence of interaction between tone and intonation in Mandarin Chinese, *J. of Phonetics*, 11, 337–351, 1983.
- Córdoba, R., J. M. Montero, J. M. Gutiérrez, J. A. Vallejo, E. Enriquez, and J. M. Pardo, Selection of most significant parameters for duration modelling in a Spanish text-to-speech system using neural networks, *Computer Speech & Language*, 16, 183–203, 2002.
- Costa, A., A. Colome, and A. Caramazza, Lexical access in speech production: the bilingual case, *Psicológica*, 21, 403–437, 2000.
- Courtenay, K., Yorùbá: A terraced-level language with three tonemes, *Studies in African Linguistics*, 2, 239–255, 1971.
- Crystal, D., *Prosodic Systems and Intonation in English*, Cambridge University Press, Cambridge, 1969.
- d'Alessandro, C., and P. Mertens, Automatic pitch contour stylization using a model of tonal perception, *Computer Speech & Language*, 9, 257–288, 1995.
- De Mori, R., R. Gubrynowicz, and P. Laface, Inference of a knowledge source for the recognition of nasals in continuous speech, *IEEE Trans. on Speech & Audio Processing*, ASSP-27, 538–549, 1979.
- Déchaine, R.-M., On the left edge of Yorùbá in complements, *Lingua*, 111, 81–130, 2001.
- Delmonte, R., SLIM prosodic automatic tools for self-learning instruction, *Speech Communication*, 30, 145–166, 2000.
- Demichelis, P., R. De Mori, P. Laface, and M. O'Kane, Computer recognition of plosive sounds using contextual information, *IEEE Trans. on Acoustics, Speech, & Signal Processing*, ASSP-31, 359–377, 1983.
- Dilley, L. C., D. R. Ladd, and A. Schepman, Alignment of L and H tone in bitonal pitch accents: testing two hypotheses, *J. of Phonetics*, 33, 115–119, 2005.
- D'Imperio, M., Language-specific and universal constraints on tonal alignment: the nature of targets and “anchors”, [http://www.isca-speech.org/archive/sp2002/sp02\\_101.html](http://www.isca-speech.org/archive/sp2002/sp02_101.html), 2002, visited: May 2005.
- Divay, M., and A. J. Vitale, Algorithms for grapheme-phoneme translation for English and French: application of database search and speech synthesis, *Computational Linguistics*, 23, 495–523, 1997.
- Dong, M., and R. Kothari, Look-ahead based fuzzy decision tree induction, *IEEE Trans. on Fuzzy Systems*, 9, 461–468, 2001.
- Donovan, R. E., Trainable Speech Synthesis, Ph.D. thesis, Cambridge University Engineering Department, Cambridge, 1996.



## BIBLIOGRAPHY

- Donovan, R. E., Topics in decision tree based speech synthesis, *Computer Speech & Language*, 17, 43–67, 2003.
- Dunning, T., Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, 19, 61–74, 1993.
- Dusterhoff, K., Synthesizing fundamental frequency using models automatically trained from data, Ph.D. thesis, University of Edinburgh, Edinburgh, 2000.
- Dusterhoff, K., and A. Black, Generating  $f_0$  contours for speech synthesis using the Tilt intonation theory, in *ESCA Workshop on Intonation Theory, Model and Application*, edited by A. Bontinis, pp. 107–110, Athens, 1997.
- Dusterhoff, K. E., A. W. Black, and P. Taylor, Using decision trees within the Tilt intonation model to predict  $f_0$  contours, in *6<sup>th</sup> European Conference on Speech Communication and Technology (EuroSpeech '99)*, pp. 1627–1630, Budapest, 1999.
- Dutoit, T., and H. Leich, MBR-PSOLA: Text-to-speech synthesis based on an MRE re-synthesis of the segments database, *Speech Communication*, 13, 435–440, 1993.
- Dutoit, T., and H. Leich, On the ability of various speech models to smooth segment discontinuities in the context of text-to-speech synthesis by concatenation, in *Proc. of UUSIPCO*, vol. 1, pp. 8–12, 1994.
- Edgington, M., A. Lowry, P. Jackson, A. P. Breen, and S. Minnis, Overview of current text-to-speech techniques: Part I: text and linguistic analysis, *BT Tech. Journal*, 14, 68–83, 1996a.
- Edgington, M., A. Lowry, P. Jackson, A. P. Breen, and S. Minnis, Overview of current text-to-speech techniques: Part II: prosody and speech generation, *BT Tech. Journal*, 14, 84–99, 1996b.
- Ehrich, R. W., and J. P. Foith, Representation of random waveforms by relational trees, *IEEE Trans. on Computers*, C-25, 725–736, 1976.
- El-Imama, Y. A., An unrestricted vocabulary Arabic speech synthesis system, *IEEE Trans. on Speech & Audio Processing*, 37, 1829–1845, 1989.
- El-Imama, Y. A., Speech synthesis using partial syllables, *Computer Speech & Language*, 4, 203–229, 1990.
- Eric, L., and M. Tatham, Word and syllable concatenation in text-to-speech synthesis, in *Proc. of European conference on Speech Communication and Technology*, vol. 2, pp. 615–618, Budapest, 1999.
- Fackrell, J. W. A., H. Vereecken, J. P. Martens, and B. V. Coile, Multilingual prosody modelling using cascades of regression trees and neural networks, <http://chardonay.elis.rug.ac.be/papers/1999-0001.pdf>, 1999, visited: Sep 2004.
- Fagyal, Z., Temporal template for background information: the scaling of pitch in utterance-medial parenthetical in French, An International Conference on Speech Prosody 2002 [URL=www.lpl.univ-aix.fr/sp2002/pdf/fagyal.pdf](http://www.lpl.univ-aix.fr/sp2002/pdf/fagyal.pdf), 2002, visited: Aug 2004.

## BIBLIOGRAPHY

- Fernanda, F., Prosody, in *Encyclopedia of Cognitive Science*, Macmillan Reference Ltd, 2000.
- Fisher, W. M., G. R. Doddington, and K. M. Goudie-Marshall, The DARPA speech recognition research database: specifications and status, in *DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.
- Flanagan, J. L., *Speech Analysis, Synthesis, and Perception*, Springer-Verlag, Berlin, 1972.
- Flanagan, J. L., and K. Ishizaka, Automatic generation of voiceless excitation in a vocal-cord tract speech synthesizer, *IEEE Trans. on Speech & Audio Processing*, ASSP-24, 163–170, 1976.
- Fletcher, J., and A. McVeigh, Segment and syllable duration in Australian English, *Speech Communication*, 13, 355–365, 1993.
- Fujisaki, H., and K. Hirose, Modelling the dynamic characteristics of voiced fundamental frequency with application to analysis and synthesis of intonation, in *13<sup>th</sup> International Congress of Linguistics*, pp. 57–70, 1982.
- Fujisaki, H., and K. Hirose, Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *Journal of the Acoustical Society of America*, 5, 1984.
- Fujisaki, H., S. Ohno, and C. Wang, A command-response model for  $f_0$  contour generation in multi-lingual speech synthesis, in *Proc. of the 3<sup>rd</sup> ESCA/COCOSDA Intl. Workshop on Speech Synthesis*, pp. 299–309, 1998.
- Fujisaki, H., R. Tomana, S. Narusawa, S. Ohno, and C. Wang, Physiological mechanism for fundamental frequency control in Standard Chinese, in *Proc. ICSLP 2000*, 2000.
- Fujisaki, H., S. Ohno, and W. Gu, Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of their  $f_0$  contours, in *International Symposium on Tonal Aspects of Languages*, Beijing, 2004, visited: Jun 2005.
- Fujisaki, H., C. Wang, S. Ohno, and W. Gu, Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model, *Speech Communication*, 47, 59–70, 2005.
- Gandour, J., S. Potisuk, and S. Dechnonhkit, Tonal co-articulation in Thai, *Phonetica*, 56, 123–134, 1999.
- Gårding, E., A generative model of intonation, in *Prosody: models and measurements*, edited by A. Cutler and D. R. Ladd, pp. 11–25, Springer, Berlin, 1983.
- Gibbon, D., Computational phonology and typology of West African tone systems, in *Typology of African Tone System*, pp. 1–8, Workshop at Universität Bielefeld, Germany, 2001.

## BIBLIOGRAPHY

- Gibbon, D., Finite state prosodic analysis of African corpus resources, in *Proc. of EuroSpeech*, 2003a.
- Gibbon, D., Finite state processing of tone languages, in *Proc. of EACL*, vol. 3, pp. 291–297, Copenhagen, 2003b.
- Gibbon, D., Tone and timing: two problems and two methods for prosody typology, in *Intl. Symposium on Tonal Aspect of Languages*, Beijing, China, 2004a.
- Gibbon, D., Tone modelling, personal email communication, 2004b.
- Goldfarb, C. F., E. J. Mosher, and T. I. Peterson, An online system for integrated text processing, in *Proc. American Society for Information Science*, vol. 7, pp. 147–150, 1970.
- Goldsmith, J., *Autosegmental and metrical phonology*, Blackwell, Oxford, 1990.
- Goldsmith, J., Dealing with prosody in a TTS system, *International Journal of Speech Technology*, 3, 51–63, 1999.
- Goubanova, O., Predicting segmental duration using Bayesian belief networks, <http://www.ssw4.org/papers/139.pdf>, 2002, visited: Mar 2005.
- Goubanova, O., and P. Taylor, Using Bayesian belief networks for model duration in text-to-speech systems, in *Proceedings of ICSLP2000*, 2000.
- Greenberg, S., Speaking in shorthand- a syllable centric perspective of understanding pronunciation variation, <http://www.citeseer.ist.psu.edu/greenberg98speaking.html>, 1998, visited: Apr 2005.
- Grice, M., M. Reyelt, R. Benzmler, J. Mayer, and A. Batliner, Consistency in transcription and labelling of German intonation with G-ToBI, in *Proc. ICSLP 96*, Philadelphia, 1996.
- Grønnum, N., The groundworks of Danish intonation: an introduction, Ph.D. thesis, University of Copenhagen/Museum Tusulanum Press, Copenhagen, 1992.
- Harrison, P., Acquiring the phonology of lexical tone in infants, *Lingua*, 110, 581–616, 2000.
- Hawkins, S., S. Heid, J. House, and M. Huckvale, Assessment of naturalness in the *Prosynth* speech synthesis project, in *IEE Colloquium on Speech Synthesis*, London, 2000.
- Hayes, B., Compensatory lengthening in moraic phonology, *Linguistic Enquiry*, pp. 253–306, 1989.
- Hendessi, F., A. Ghayoori, and T. A. Gulliver, A speech synthesizer for Persian text using a neural network with a smooth ergodic HMM, [www.ece.uvic.ca/agul-live/FinaljournalPaper2.pdf](http://www.ece.uvic.ca/agul-live/FinaljournalPaper2.pdf), 2002.
- Hermes, D. J., Measuring the perceptual similarity of pitch contour, *Journal of Speech Language and Hearing Research*, 41, 73–82, 1998.

## BIBLIOGRAPHY

- Hertz, S. R., and M. K. Huffman, A nucleus-based timing model applied to multi-dialect speech synthesis by rule, in *International Conference on Spoken Language Processing*, vol. 2, pp. 1171–1174, 1992.
- Hertz, S. R., and L. Zsiga, The Delta system with SYLLT: increases capability for teaching and research in phonetics, in *International Conference on Spoken Language Processing*, vol. 2, pp. 322–325, 1995.
- Hirst, D., Prediction of prosody: an overview, in *Talking Machines: Theories, Models, and Designs*, edited by G. Bailly, C. Benoit, and T. R. Sawallis, pp. 77–82, Elsevier Science, Amsterdam, 1992.
- Hirst, D. J., A. D. Cristo, and R. Espesser, Levels of representation and levels of analysis for intonation, in *Theory and Experiment Studies presented to Gösta Bruce*, edited by M. Horne, Kluwer, Dordrecht, 2000.
- Höhne, H. D., C. Coker, S. E. Levinson, and L. R. Rabiner, On the temporal alignment of sentence of natural and synthetic speech, *IEEE Trans. on Speech & Audio Processing*, *ASPP-31*, 807–813, 1983.
- Holmes, W. J., and M. J. Russell, Probabilistic trajectory segmental HMMs, *Computer Speech & Language*, *13*, 3–37, 1999.
- Hombert, J.-M., Perception of tones of bisyllabic nouns in Yorùbá, *Studies in African Linguistics*, *Suppl. 6*, 109–121, 1976.
- Hombert, J.-M., Consonant types, vowel height and tones in Yorùbá, *Studies in African Linguistics*, *8*, 109–121, 1977.
- Hombert, J.-M., Consonant types, vowel quality, and tone, in *Tone: A Linguistic survey*, edited by V. Fromkin, pp. 77–111, Academic Press, New York, 1978.
- Huang, H.-P., and C.-C. Liang, Strategy-based decision making of a soccer robot system using a real-time self-organising fuzzy decision tree, *Fuzzy Sets & Systems*, *127*, 49–64, 2002.
- Huang, X., A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, Recent improvements on Microsofts trainable text-to-speech system- Whistler, in *IEEE Proc. ICASSP.*, pp. 959–962, Munich, 1997.
- Huang, X. D., Y. Ariki, and M. Jack, *Hidden Markov Models for speech recognition*, Information Technology Series, Edinburgh University Press, Edinburgh, 1990.
- Huckvale, M., 10 things engineers have discovered about speech recognition, in *NATO ASI workshop on speech pattern processing*, 1997.
- Huckvale, M., Representation and processing of linguistic structures for an all-prosodic synthesis system using XML, in *Proc. of EuroSpeech '99*, 4, pp. 1847–1850, Budapest, 1999.

## BIBLIOGRAPHY

- Huckvale, M., The use and potential of extensible markup (XML) in speech generation, in *Improvements in Speech Synthesis: Cost 258: The Naturalness of Synthetic Speech*, edited by E. Keder, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale, chap. 30, pp. 298–305, Wiley Inter. Science, 2001.
- Huckvale, M., Speech synthesis, speech simulation and speech science, in *Proc. International Conference on Speech and Language Processing*, pp. 1261–1264, Denver, 2002.
- Huggins, A. W. F., and R. S. Nickerson, Speech quality evaluation using “phoneme-specific” sentences, *Journal of the Acoustical Society of America*, 77, 1896–1906, 1985.
- Hunt, A. J., and A. W. Black, Unit selection in a concatenative speech synthesis system using a large database, in *Proc. Of IEEE International Conf. on Acoustic and Speech Signal Processing*, vol. 1, pp. 375–376, 1996.
- Iida, A., A study on corpus-based speech synthesis with emotion, Ph.D. thesis, Graduate School of Media and Governance, Keio University, Japan, 2002.
- Inverson, G. K., and D. W. Wheeler, Phonological categories and constituents, in *Linguistic Categorisation*, pp. 93–114, Benjamins, Amsterdam, 1989.
- Isard, S., and D. Miller, Diphone synthesis techniques, in *Proc. Of IEE Speech Input/Output Conference*, pp. 77–82, IEE, 1986.
- Iwahashi, N., and Y. Sagisaka, Statistical modelling of speech segment duration by constrained tree regression, *IEICE TRANS. INF. & SYST*, E83D, 1550–1559, 2000.
- Jang, J., Structure determination in fuzzy modelling: A fuzzy CART approach, in *IEEE Conf. Fuzzy Systems*, pp. 480–485, 1994.
- Janikow, C. Z., Fuzzy processing in decision trees, in *Proc. of the 6<sup>th</sup> Int. Symp. on Artificial Intelligence*, pp. 360–367, 1993.
- Janikow, C. Z., Fuzzy decision trees: issues and methods, *IEEE Trans. on Systems, Man, & Cybernetics*, 28, 1–14, 1998.
- Janikow, C. Z., FID33 fuzzy decision tree,  
<http://www.cs.umsl.edu/~janikow/fid/fid32/overview.htm>, 2004, visited: Jan 2005.
- JavaSpeechML, Java Speech Markup Language (JSML) Specification , Version 0.5, 1997, visited: May 2005.
- Jensen, U., R. K. Moore, P. Dalsgaard, and B. Lindberg, Modelling intonation contour at the phrase level using continuous density Hidden Markov Models, *Computer Speech & Language*, 8, 247–260, 1994.
- Jitca, D., H. N. Teodorescu, V. Apopei, and F. Grigoras, Improved speech synthesis using fuzzy methods, *International Journal of Speech Technology*, 5, 227–235, 2002.

## BIBLIOGRAPHY

- Jun, S., and M. Oh, A prosodic analysis of three types of wh-phrase in Korean, *Language and Speech*, 39, 37–61, 1996.
- Kaiki, N., K. Takeda, and Y. Sagisaka, Statistical duration rules in Japanese speech synthesis, in *Proc. of ICSLP*, vol. 1, pp. 17–20, Springer-Verlag, Kobe, 1990.
- Keller, E., and B. Zellner, A statistical timing model for French, in *XIIIème Cong. Int. Des. Sci. Phon.*, vol. 3, pp. 302–305, Stockholm, 1995.
- Kelly, J., and L. Gerstman, An artificial talker driven from a phonetic input (abstract), *Journal of the Acoustical Society of America*, 33, 835, 1961.
- Kelly, J., and J. Local, Long-domain resonance pattern in English, in *Proc. of IEEE Speech Input/Output Conference*, vol. 258, pp. 77–82, 1986.
- Kessler, B., and R. Treiman, Syllable structure and the distribution of phonemes in English syllables, *Journal of Memory and Language*, 37, 295–311, 1997.
- Kiat-arpakul, R., J. Fakcharoenphol, and S. Keretho, A combined phoneme-based and demissyllable-based approach for Thai speech synthesis, in *Proc. of 2<sup>nd</sup> Symposium on Natural Language Processing*, pp. 361–369, 1995.
- King, S., and R. A. J. Clark, Use of Wagon in duration modelling, Personal communication, CSTR, Edinburgh, 2004.
- Klatt, D. H., Linguistic uses of segmental duration in English: acoustic and perceptual evidence, *Journal of the Acoustical Society of America*, 59, 1208–1221, 1976.
- Klatt, D. H., Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America*, 82, 737–793, 1987.
- Klatt, D. H., and L. C. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America*, 87, 820–857, 1990.
- Kochanski, G., and C. Shih, Prosody modelling with soft templates, *Speech Communication*, 39, 311–352, 2003.
- Kochanski, G., C. Shih, and H. Jing, Quantitative measurement of prosody strength in Mandarin, *Speech Communication*, 41, 625–645, 2003a.
- Kochanski, G., C. Shih, and H. Jing, Hierarchical structure and word strength prediction of Mandarin prosody, *International Journal of Speech Technology*, 6, 33–43, 2003b.
- Kohler, K. J., Invariability and variability in speech timing: from utterance to segment in German, in *Invariability and Variability in Speech Processes*, edited by J. Perkell and D. H. Klatt, pp. 268–298, Hillsdale: Erlbaum, 1986.
- Kohler, K. J., Modelling prosody in spontaneous speech, in *Computing Prosody*, edited by Y. Sagisaka, N. Campbell, and N. Higuchi, pp. 187–210, Springer, New York, 1997a.

## BIBLIOGRAPHY

- Kohler, K. J., Parametric control of prosodic variables by symbolic input in TTS synthesis, in *Progress in Speech Synthesis*, edited by J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, pp. 459–475, Springer, New York, 1997b.
- Kornai, A., Relating phonetic and phonological categories, *Discrete Mathematics and Theoretical Computer Science*, pp. 1–15, 1993.
- Kosanović, B. R., L. F. Chaparro, and R. J. Sciabassi, Signal analysis in fuzzy information space, *IEEE Trans. on Fuzzy Systems*, 77, 49–62, 1996.
- Kosko, B., Fuzzy systems as universal approximators, *IEEE Trans. on Computers*, 43, 1329–1333, 1994.
- Ladd, D. R., A model of intonation phonology for use in speech synthesis by rule, in *European Conference on Speech Technology (ESCA)*, pp. 21–24, 1987.
- Ladd, D. R., Metrical representation of pitch register, in *Papers in Laboratory Phonology 1: Between the Grammar and the Physics of Speech*, edited by J. Lingston and M. Beckman, pp. 35–57, Cambridge University Press, Cambridge, 1990.
- Ladd, D. R., *Intonational Phonology*, vol. 79 of *Cambridge Studies in Linguistics*, Cambridge University Press, Cambridge, 1996.
- Ladd, D. R., Tones and turning points: Bruce, Pierrehumbert, and the elements of intonation phonology, in *Prosody: Theory and Experiment - Studies presented to Gösta Bruce*, edited by M. Horne, pp. 37–50, Kluwer Academic Publishers, Dordrecht, 2000.
- Láníran, Y. O., and G. N. Clements, A long-distance dependency in Yorùbá tone realization, in *Proc. of XIII<sup>th</sup> International Congress of Phonetic Science*, vol. 2, pp. 743–737, Stockholm, 1995.
- Láníran, Y. O., and G. N. Clements, Downstep and high rising: interacting factors in Yorùbá tone production, *J. of Phonetics*, 31, 203–250, 2003.
- Leben, W., The tones in English intonation, *Linguistic Analysis*, 2, 69–107, 1976.
- Lee, C. C., Fuzzy logic in control systems: fuzzy logic controller, Part I, *IEEE Trans. on Fuzzy Systems*, 20, 404–418, 1990a.
- Lee, C. C., Fuzzy logic in control systems: fuzzy logic controller, Part II, *IEEE Trans. on Fuzzy Systems*, 20, 419–435, 1990b.
- Lee, K.-S., and R. V. Cox, A segmental speech coder based on a concatenative TTS, *Computer Speech & Language*, 38, 89–100, 2002.
- Lee, L.-S., C.-Y. Tseng, and M. Ouh-Young, The synthesis rules in a Chinese text-to-speech system, *IEEE Trans. on Speech & Audio Processing*, 37, 1309–1320, 1989.
- Lee, L.-S., C.-Y. Tseng, and C.-J. Hsieh, Improved tone concatenation rules in a formant-base Chinese text-to-speech systems, *IEEE Trans. on Speech & Audio Processing*, 1, 287–294, 1993.

## BIBLIOGRAPHY

- Lee, S., and Y.-H. Oh, Tree-based modelling of prosodic phrasing and segmental duration for Korean TTS systems, *Speech Communication*, 28, 283–300, 1999.
- Lee, S. C., S. Xu, and B. Guo, Microcomputer-generated Chinese speech, *Computer Processing of Chinese & Oriental Languages*, 1, 87–103, 1983.
- Lee, T., G. Kochanski, C. Shih, and Y. Li, Modeling tones in continuous Cantonese speech, in *Proc. of Int. Conf. on Spoken Language Processing*, Denver, Colorado, 2002a, visited: Apr 2005.
- Lee, T., W. Lau, Y. W. Wong, and P. C. Ching, Using tone information in Cantonese continuous speech recognition, *ACM Trans. on Asian Language Information Processing*, 1, 83–102, 2002b.
- Lee, W.-S., The effect of intonation on the citation tones in Cantonese, in *International Symposium on Tonal Aspect of Language*, pp. 28–31, Beijing, 2004.
- Lemmetty, S., Review of Speech Synthesis Technology, Master's thesis, Department of Electrical and Communication Engineering, Helsinki University, Helsinki, 1999.
- Levinson, S. E., Continuously variable duration Hidden Markov Models for speech analysis, in *Proc. of IEEE ICASSP*, pp. 1241–1244, 1986.
- Levitt, H., and L. R. Rabiner, Analysis of fundamental frequency contours in speech, *Journal of the Acoustical Society of America*, 50, 637–655, 1971.
- Lewis, E., and M. Tatham, Word and syllable concatenation in text-to-speech synthesis, in *6<sup>th</sup> European Conference on Speech Communications and Technology*, pp. 615–618, 1999.
- Li, P.-Y., Perceptual analysis of the six contrastive tones in Cantonese, in *International Symposium on Tonal Aspects of Languages*, pp. 119–122, Beijing, 2004, visited: Jun 2005.
- Li, Y., T. Lee, and T. Qian, Analysis and modelling of  $f_0$  contours for Cantonese text-to-speech, *ACM Trans. on Asian Language Information Processing*, 3, 169–180, 2004.
- Liang, N. Y., and Y. B. Zhen, A Chinese word segmentation model and a Chinese word segmentation system, in *PS-CWW, COLIP*, vol. 1, pp. 51–55, 1991.
- Liberman, A. M., The grammars of speech and language, *Cognitive Psychology*, 1, 301–323, 1970.
- Liberman, A. M., and J. Pierrehumbert, Intonational invariant under changes in pitch range and length, in *Language and Sound Structure*, edited by M. Aronoff and R. T. Oehrle, pp. 157–233, MIT Press, Cambridge, MA, 1984.
- Liberman, A. M., and R. Sproat, The stress and structure of modified noun phrase in English, in *Lexical Matters*, edited by I. Sag, University of Chicago Press, Chicago, 1992.



## BIBLIOGRAPHY

- Liberman, M., Computer speech synthesis: its status and prospects, in *National Academy of Science*, vol. 92, pp. 9928–9931, Irvine, 1995.
- Liberman, M., and A. Prince, On stress and linguistic rhythm, *Linguistic Inquiry*, 8, 249–336, 1977.
- Lin, C.-H., R.-C. Wu, J.-Y. Chang, and S.-F. Liang, A novel prosodic-information synthesizer based on recurrent fuzzy neural networks for Chinese TTS system, *IEEE Trans. on Systems, Man, & Cybernetics, B*, 1–16, 2003.
- Lindau, M., Testing a model of intonation in a tone language, *Journal of the Acoustical Society of America*, 80, 757–764, 1986.
- Lindblom, B., and Q. Engstrand, In what sense is speech quantal, *Journal of Phonetics*, 17, 107–121, 1989.
- Ljolje, A., and F. Fallside, Synthesis of natural sounding pitch contour in isolated utterances using Hidden Markov Models, *IEEE Trans. on Speech & Audio Processing, ASSP-34*, 1074–1080, 1986.
- Lo, W. K., T. Lee, and P. C. Ching, Development of Cantonese spoken language corpora for speech applications, in *Proc. Of ISCSPL-98*, pp. 102–107, Singapore, 1998.
- Local, J. K., Modelling assimilation in non-segmental, rule-free synthesis, in *Papers in Laboratory Phonology II: gesture, segment, prosody*, edited by G. J. Docherty and D. R. Ladd, pp. 190–223, Cambridge University Press, Cambridge, 1992.
- Louw, J. A., and E. Barnard, Automatic intonation modelling with INTSINT, <http://www.csir.co.za/websource/ptl0002/docs/HLT/louwja04intsint.pdf>, 2002, visited: Jun 2004.
- Lu, J., N. Umeni, G. Li, and T. Ifikube, Tone enhancement in Mandarin speech for listeners with hearing impairment, *IEICE Tran. Inf. & Syst., E84-D*, 651–661, 2001.
- Maddieson, I., *Patterns of sounds*, 2<sup>nd</sup> ed., Cambridge University Press, 1984.
- Manteescu, D., English phonetics and phonological theory, 20<sup>th</sup> century approach, <http://www.unibuc.ro/eBooks/filologie/mateescu/pdf/76.pdf>, 2004, visited: Aug 2004.
- Matousek, J., D. Tihelka, J. Psutka, and J. Hesová, *German and Czech Speech Synthesis Using HMM-Based Speech Segment Database*, vol. 2448 of *Lecture Notes In Computer Science*, pp. 173–180, Springer-Verlag, London, 2002.
- Mertens, P., The Prosogram: semi-automatic transcription of prosody based on a tonal perception model, <http://bach.arts.kuleuven.be/pmertens/papers/sp2004.pdf>, 2004, visited: Dec 2004.
- Mitaim, S., and B. Kosko, The shape of fuzzy sets in adaptive function approximation, *IEEE Trans. on Fuzzy Systems*, 9, 637–656, 2001.

## BIBLIOGRAPHY

- Mitra, S., K. M. Konwar, and S. K. Pal, Fuzzy decision tree, linguistic rules and fuzzy knowledge-based network: generation and evaluation, *IEEE Trans. on Systems, Man, & Cybernetics*, 32, 328–339, 2002.
- Mittrapiyanuruk, P., C. Hansakunbuntheung, V. Tesprasit, and V. Sornlertlamvanich, Improving naturalness of Thai text-to-speech synthesis by prosodic rule, in *ICSLP-2000*, vol. 3, pp. 334–337, 2000.
- Mixdorff, H., A novel approach to fully automatic extraction of Fujisaki model parameters, in *Proc. ICASSP 2000*, vol. 3, pp. 1281–1284, Istanbul, 2000.
- Möbius, B., Rare events and closed domains: two delicate concepts in speech synthesis, *International Journal of Speech Technology*, 6, 57–71, 2003.
- Möbius, B., and J. P. H. van Santen, Modelling segmental duration in German text-to-speech synthesis, in *International Conference on Spoken Languages Processing (ICSLP)*, pp. 2395–2398, 1996.
- Möbius, B., M. Pätzold, and W. Hess, Analysis and synthesis of German  $f_0$  contours by means of Fujisaki's model, *Speech Communication*, pp. 53–61, 1993.
- Monaghan, A., Markup for speech synthesis: a review and some suggestions, in *Improvements in Speech Synthesis: Cost 258: The Naturalness of Synthetic speech*, edited by E. Keder, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale, chap. 31, pp. 307–319, Wiley Inter. Science, 2001.
- Monaghan, A. I. C., Rhythm and stress shift in speech synthesis, *Computer Speech & Language*, 4, 71–78, 1990.
- Monaghan, A. I. C., The intonation of textual anomalies in text-to-speech, *Speech Communication*, 12, 371–382, 1993.
- Monaghan, A. I. C., A metrical model of prosody for multilingual TTS, *International Journal of Speech Technology*, 4, 73–81, 2003.
- Monaghan, A. I. C., and D. R. Ladd, Symbolic output as the basis for evaluating intonation in text-to-speech synthesis system, *Speech Communication*, 9, 305–314, 1990.
- Mori, R. D., P. Laface, and Y. Mong, Parallel algorithms for syllable recognition in continuous speech, *IEEE Trans. on Pattern Analysis & Machine Intelligence, PAMI-7*, 56–69, 1985.
- Morlec, Y., G. Bailly, and V. Aubergé, Generating prosodic attitudes in French: data, model and evaluation, *Speech Communication*, 33, 357–371, 2001.
- Moulines, E., and F. Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication*, 9, 453–467, 1990.

## BIBLIOGRAPHY

- Mullennix, J. W., S. E. Stern, S. J. Wilson, and C.-I. Dyson, Social perception of male and female computer synthetic speech, *Computers in Human Behaviour*, 19, 407–242, 2003.
- Nie, J. Y., X. Ren, and M. Brisebois, A unifying approach to segmentation of Chinese and its application to text retrieval, in *Proc. of Research on Computational Linguistics Conference (ROCLING VIII)*, pp. 175–190, Taiwan, 1995.
- O'Brien, D., and A. I. C. Monaghan, Concatenative synthesis based on a harmonic model, *IEEE Trans. on Speech & Audio Processing*, 9, 11–19, 2001.
- O'Brien, S. M., Spectral features of plosives in connected-speech signals, *Int. J. Man-Machine Studies*, 38, 97–127, 1993.
- Odéjobí, O. A., A. J. Beaumont, and S. H. S. Wong, Experiments on stylisation of standard Yorùbá language tones, *Tech. Rep. KEG/2004/003*, Aston University, Birmingham, 2004a.
- Odéjobí, O. A., A. J. Beaumont, and S. H. S. Wong, A computational model of intonation for Yorùbá text-to-speech synthesis: design and analysis, in *Lecture Notes in Artificial Intelligence*, edited by P. Sojka, I. Kopeček, and K. Pala, Lecture Notes in Computer Science (LNAI 3206), pp. 409–416, Springer-Verlag, Berlin, 2004b.
- Ogden, R., J. Local, and P. Carter, Temporal interpretation in PROSYNTH, a prosodic speech synthesis system, [http://www-users.york.ac.uk/~lang19/papers/york\\_icphs99.pdf](http://www-users.york.ac.uk/~lang19/papers/york_icphs99.pdf), 1999, visited: May 2005.
- Ogden, R., S. Hawkins, J. House, M. Huckvale, J. Local, P. Carter, J. Dankovicova, and S. Heid, ProSynth: an integrated prosodic approach to device-independent natural-sounding speech synthesis, *Computer Speech & Language*, 14, 177–210, 2000.
- Ògúnbòwálé, P. O., *Àsà Ìbílẹ̀ Yorùbá*, University Press Limited, Jericho, Ibadan, Nigeria, 1966.
- Ògúnbòwálé, P. O., *The Essentials of the Yorùbá Language*, Hodder and Stoughton, London, 1970.
- Olaru, C., and L. Wehenkel, A complete fuzzy decision tree technique, *Fuzzy Sets & Systems*, 138, 221–254, 2003.
- Olaszy, G., and G. Németh, Prosody generation for German CTS/TTS systems (from theoretical intonation pattern to practical realisation), *Speech Communication*, 21, 37–60, 1997.
- Olive, J. P., and L. H. Nakatani, Rule-synthesis of speech by word concatenation: a first step, *Journal of the Acoustical Society of America*, 55, 660–666, 1974.
- Oliver, D., Deriving pitch accent classes using automatic  $f_0$  stylisation and unsupervised clustering techniques, in *Proceedings of 2<sup>nd</sup> Baltic Conference on Human Language Technologies*, pp. 161–166, Tallinn, 2005.

## BIBLIOGRAPHY

- O'Shaughnessy, D., and J. Allen, Linguistic modality effects on fundamental frequency in speech, *Journal of the Acoustical Society of America*, 74, 1155–1171, 1983.
- Owólabí, K., *Ìjìnlẹ̀ Ìtupalẹ̀ èdè Yorùbá:Fònètíàkì àti Fonólójì*, vol. 1, 1<sup>th</sup> ed., Onibonoje Press & Book Industries (Nig.) Ltd., Ìbàdàn, 1998.
- Pal, S. K., and D. D. Majumder, Fuzzy sets and decision making approaches in vowel and speaker recognition, *IEEE Trans. on Systems, Man, & Cybernetics*, SMC-7, 625–629, 1977.
- Parthasarathy, S., and C. H. Coker, On automatic estimation of articulatory parameters in a text-to-speech system, *Computer Speech & Language*, 6, 37–75, 1992.
- Passonneau, R., and D. Litman, Empirical analysis of three dimensions of spoken discourse: segmentation, coherence, and linguistic devices, in *Computational and Conversational Discourse: Burning Issues-an Interdisciplinary Account*, edited by E. Hovey and D. Scott, pp. 161–194, Springer, Berlin, 1996.
- Pedrycz, W., and Z. A. Sosnowski, The design of decision trees in the framework of granular data and their application to software quality models, *Fuzzy Sets & Systems*, 123, 271–290, 2001.
- Pellegino, F., and R. Andre-Obrencht, Automatic language identification: an alternative approach to phonetic modelling, *Signal Processing*, 80, 1231–1244, 2000.
- Petrucelli, J. D., B. Naudrum, and M. Chen, *Applied Statistics for Engineers and Scientists*, Prentice Hall, New Jersey, 1999.
- Pierrehumbert, J., Synthesizing intonation, *Journal of the Acoustical Society of America*, 70, 985–995, 1981.
- Pierrehumbert, J., Tonal elements and their alignment, in *Prosody: Theory and Experiment. Studies presented to Gösta Bruce*, edited by M. Horne, Text, Speech and Language Technology, pp. 11–26, Kluwer, Dordrecht, 2000.
- Pirker, H., K. Alter, E. Rank, J. Matiassek, H. Trost, and G. Kubin, A system of stylized intonation contour in German, in *ESCA, EuroSpeech '97*, pp. 307–311, Rhodes, Greece, 1997.
- Pitrelli, J., M. Beckman, and J. Hirschberg, Evaluation of prosody transcription labelling reliability in the ToBI framework, in *Proc. of the 3<sup>rd</sup> Intl. Conf. on Spoken Language Processing (ICSPL'94)*, vol. 2, pp. 135–151, 1994.
- Plumpe, M., and S. Meredith, Which is more important in a concatenative text-to-speech system pitch, duration or spectral discontinuity?, in *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 25–29, Jenolan, Australia, 1998.
- Prevost, S., and M. Steedman, Specifying intonation from context for speech synthesis, *Speech Communication*, 15, 139–153, 1994.

## BIBLIOGRAPHY

- Quené, H., and R. Kager, The derivation of prosody for text-to-speech from prosodic sentence structure, *Computer Speech & Language*, 6, 77-98, 1992.
- Quinlan, J. R., Induction of decision trees, *Machine Learning*, 1, 81-106, 1986.
- Quinlan, J. R., Decision trees and decision making, *IEEE Trans. on Systems, Man, & Cybernetics*, 20, 339-346, 1990.
- Rahim, M., C. Goodyear, W. Kleijn, J. Schroeter, and M. Sondhi, On the use of neural networks in articulatory speech synthesis, *Journal of the Acoustical Society of America*, 93, 1109-1121, 1993.
- Rana, D. S., G. Hurst, L. Shepstone, J. Pilling, J. Cockburn, and M. Crawford, Voice recognition for radiology reporting: is it good enough?, *Clinical Radiology*, 60, 1205-1212, 2005.
- Raptis, S., and G. V. Carayannis, Fuzzy logic for rule-based formant speech synthesis, in *EuroSpeech '97*, pp. 1599-1602, 1997.
- Riley, M. D., Tree-based modelling of segmental durations, in *Talking Machines: Theories, Models and Designs*, edited by G. Bailly, C. Benoit, and T. R. Sawallis, pp. 265-273, Elsevier, Amsterdam, 1992.
- Ross, K. N., Modelling intonation for speech synthesis, Ph.D. thesis, Boston University, Boston, 1995.
- Rubin, P. E., T. Baer, and P. Mermelstein, An articulatory synthesizer for perceptual research, *Journal of the Acoustical Society of America*, 70, 321-328, 1981.
- Safavian, S. R., and D. Landgrebe, A survey of decision tree classifier methodology, *IEEE Trans. on Systems, Man, & Cybernetics*, 21, 660-674, 1991.
- Sakurai, A., K. Hirose, and N. Minematsu, Data-driven generation of  $f_0$  contours using a superpositional model, *Speech Communication*, 40, 535-549, 2003.
- Schroeder, M. R., A brief history of synthetic speech, *Speech Communication*, 13, 231-237, 1993.
- Shaw, S. W., and R. J. P. Defigueiredo, Structural processing of waveforms as trees, *IEEE Trans. on Speech & Audio Processing*, 38, 328-338, 1990.
- Shen, X. S., M. Lin, and J. Yan,  $f_0$  turning point as an  $f_0$  cue to tonal contrast: a case study of Mandarin tones 2 and 3, *Journal of the Acoustical Society of America*, 93, 2241-2243, 1993.
- Shih, C., Study of vowel variation for Mandarin speech synthesis, in *Proc. Of EuroSpeech*, pp. 1807-1810, 1995.
- Shih, C., A declination model of Mandarin Chinese, in *Intonation: Analysis, Modelling and Technology*, edited by A. Botinis, pp. 243-268, Kluwer Academic Publishers, 2000.

## BIBLIOGRAPHY

- Shih, C., and B. Ao, Duration study for Bell Laboratories Mandarin text-to-speech system, in *Progress in Speech Synthesis*, edited by J. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, pp. 383–399, Springer, New York, 1997.
- Shih, C., and R. Sproat, Issues in text-to-speech conversion for Mandarin, *Computer Processing of Chinese & Oriental Languages*, 1, 87–103, 1983.
- Shih, C., and R. Sproat, Issues in text-to-speech conversion for Mandarin, *Computational Linguistics and Chinese Language Processing*, 1, 37–86, 1996.
- Silverman, K., and J. Pierrehumbert, The timing of prenuclear high accents in English, in *Papers in Laboratory Phonology I*, edited by J. Kingston and M. E. Beckman, pp. 72–106, Cambridge University Press, Cambridge, 1990.
- Silverman, K. E. A., The structure and processing of  $f_0$  contours, Ph.D. thesis, Cambridge University, Cambridge, 1987.
- Sison, L. G., and E. K. P. Chong, Fuzzy modelling by induction and pruning of decision trees, in *IEEE Intl. Symposium on Intelligent Control*, pp. 166–171, Columbus, Ohio, 1994.
- Sjolander, K., and J. Beskow, WaveSurfer 1.7, <http://www.speech.kth.se/wavesurfer/>, 2004, visited: Jun 2004.
- Smith, C. L., Topic transitions and durational prosody in reading aloud: production and modelling, *Speech Communication*, 42, 247–270, 2004.
- Sproat, R., C. Shih, W. Gale, and N. Chang, A stochastic finite-state word-segmentation algorithm for Chinese, *Computational Linguistics*, 22, 377–404, 1996.
- Sproat, R., A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo, and M. Edgington, SABLE: A standard for TTS markup, <http://www.bell-labs.com/project/tts/sabpap/sabpap.html>, 1998, visited: Jun 2004.
- Stewart, J. M., Dschang and Ebrié as Akan-type tonal downstep languages, in *The phonology of tone: the representation of tonal register*, edited by H. van der Hulst and K. Smith, pp. 185–244, Mouton de Gruyter, Berlin, 1993.
- Suárez, A., and J. F. Lutsko, Globally optimal fuzzy decision trees for classification and regression, *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 21, 1297–1311, 1999.
- Sugeno, M., and T. Yasukawa, A fuzzy-logic-based approach to qualitative modelling, *IEEE Trans. on Fuzzy Systems*, 1, 7–31, 1993.
- Sun, X., The determination, analysis, and synthesis of fundamental frequency, Ph.D. thesis, Northwestern University, Evanston, IL, 2002.
- Syrdal, A. K., J. Hirschberg, J. McGory, and M. Beckman, Automatic ToBI prediction and alignment to speech manual labelling of prosody, *Speech Communication*, 33, 135–151, 2001.

## BIBLIOGRAPHY

- 't Hart, J., R. Collier, and A. Cohen, A perceptual study of intonation: an experimental-phonetic approach to speech melody, in *Cambridge Studies in Speech Science and Communication*, Cambridge University Press, Cambridge, 1990.
- Takagi, T., and M. Sugeno, Fuzzy identification of systems and its application to modelling and control, *IEEE Trans. on Systems, Man, & Cybernetics*, SMC-1, 116–132, 1985.
- Takano, S., K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, A Japanese TTS system based on multiform units and speech modification algorithm with Harmonics Reconstruction, *IEEE Trans. on Speech & Audio Processing*, 9, 3–10, 2001.
- Tarsaku, P., V. Sornlertlamvanivh, and R. Thongpresirt, Thai grapheme-to-phoneme using probabilistic GLR parser, in *Proc. of EuroSpeech*, vol. 2, pp. 1057–1060, 2001.
- Tatham, M., and E. Lewis, Syllable reconstruction in concatenated waveform speech synthesis, in *Proc. of International Congress of Phonetic Science*, edited by J. Ohala, pp. 2303–2306, San Francisco, 1999.
- Taylor, C., Typesetting African languages, <http://www.ideography.co.uk/library/afrolingua.html>, 2000a, visited: Apr 2004.
- Taylor, P., The rise/fall/connection model of intonation, *Speech Communication*, 15, 169–186, 1994.
- Taylor, P., Analysis and synthesis of intonation using the Tilt model, *Journal of the Acoustical Society of America*, 107, 1697–1714, 2000b.
- Taylor, P., and A. Isard, SSML: A speech synthesis markup language, *Speech Communication*, 21, 123–133, 1997.
- Taylor, P. A., A phonetic model of English intonation, Ph.D. thesis, University of Edinburgh, Edinburgh, 1992.
- 'tHart, J.,  $f_0$  stylization in speech: straight lines versus parabolas, *Journal of the Acoustical Society of America*, 6, 3368–3370, 1991.
- 'tHart, J., and A. Cohen, Intonation by rule: a perceptual quest, *J. of. Phonetics*, 1, 309–327, 1972.
- Thorsen, N. G., Intonation and text in standard Danish, *Journal of the Acoustical Society of America*, 77, 1205–1216, 1985.
- Thorsen, N. G., Sentence intonation in textual context-supplementary data, *Journal of the Acoustical Society of America*, 80, 1041–1047, 1986.
- Thubthong, N., and B. Kijirikul, Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model, *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems*, 9, 815–825, 2001.

## BIBLIOGRAPHY

- Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1315–1318, Istanbul, 2000.
- Tokuda, K., , T. Masuko, N. Miyazaki, and T. Kobayashi, Multi-space probability distribution HMM, *IEICE Trans. Inf. & Syst.*, *E85-D*, 455–464, 2002a.
- Tokuda, K., H. Zen, and A. W. Black, An HMM-based speech synthesis system applied to English, in *IEEE Speech Synthesis Workshop*, pp. 11–13, Santa Monica, California, 2002b.
- Vainio, M., Artificial neural network based prosody models for Finnish text-to-speech synthesis, Ph.D. thesis, Department of Phonetics, University of Helsinki, Helsinki, 2001.
- van Santen, J., and J. Hirschberg, Segmental effects on timing and height of pitch contours, in *Proc. of 'ICSLP'*, vol. 2, pp. 719–722, Yokohama, 1994.
- van Santen, J. P. H., Contextual effects on vowel duration, *Speech Communication*, *11*, 513–546, 1992.
- van Santen, J. P. H., Assignment of segmental duration in text-to-speech synthesis, *Computer Speech & Language*, *8*, 95–128, 1994.
- van Santen, J. P. H., and J. P. Olive, The analysis of contextual effects on segmental duration, *Computer Speech & Language*, *4*, 359–390, 1990.
- van Santen, J. P. H., C. Shih, B. Möbius, E. Tzoukermann, and M. Tanenblatt, Multilingual duration modelling, in *European Conference on Speech Communication and Technology*, pp. 2651–2655, 1997.
- Venditti, J., Japanese ToBI Labelling Guidelines, [http://www.ling.ohio-state.edu/research/phonetics/J\\_ToBI/jtobi.html/jtobi.html](http://www.ling.ohio-state.edu/research/phonetics/J_ToBI/jtobi.html/jtobi.html), 1995, visited: May 2004.
- Venditti, J. J., and J. P. H. van Santen, Modelling duration for Japanese text-to-speech synthesis, in *Proc. of the 3<sup>rd</sup> Intl. Workshop on Speech Synthesis*, pp. 31–36, Jenolan Caves, Australia, 1998.
- Véronis, J., P. D. Cristo, F. Courtios, and C. Chaumette, A stochastic model of intonation for text-to-speech synthesis, *Speech Communication*, *26*, 233–244, 1998.
- Viana, M. C., L. C. Oliveira, and A. I. Mata, Prosody phrasing: machine and human evaluation, *International Journal of Speech Technology*, *6*, 83–94, 2003.
- Viswanathan, M., and M. Viswanathan, Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale, *Computer Speech & Language*, *19*, 55–83, 2005.
- VoiceXML, Voice eXtensible Markup Language: VoiceXML, <http://www.voicexml.org/specs/VoiceXML-100.pdf>, 2000, visited: May 2003.



## BIBLIOGRAPHY

- Wang, C., Prosodic modelling for improved speech recognition and understanding, Ph.D. thesis, Massachusetts Institute of Technology, 2001.
- Wang, H., and D. Qiu, Computing with words via Turing machines: a formal approach, *IEEE Trans. on Fuzzy Systems*, 11, 742–753, 2003.
- Wang, H.-C., and T.-H. Hwang, An initial study on the speech synthesis of Taiwanese Hokkian, *Computer Processing of Chinese & Oriental Languages*, 7, 21–36, 1993.
- Wang, T.-R., and S.-H. Chen, Tone recognition of continuous Mandarin speech assisted with prosodic information, *Journal of the Acoustical Society of America*, 96, 2637–2645, 1994.
- Wang, W.-J., Y.-F. Liao, and S.-H. Chen, RNN-based prosodic modelling for Mandarin speech and its application to speech-to-text conversion, *Speech Communication*, 36, 247–265, 2002.
- Weitian, C., Sufficient conditions on fuzzy logic controllers as universal approximators, *IEEE Trans. on Systems, Man, & Cybernetics*, 31, 270–274, 2001.
- Wightman, C. W., ToBI or not ToBI, in *International Symposium on Speech Prosody*, Aix-en-Provence, 2002, visited: Jun 2004.
- Wightman, C. W., and M. Ostendorf, Automatic labeling of prosodic patterns, *IEEE Trans. on Speech & Audio Processing*, 2, 469–481, 1994.
- Wikipedia, Document Type Definition, [http://en.wikipedia.org/wiki/Document\\_Type\\_Definition](http://en.wikipedia.org/wiki/Document_Type_Definition), 2004, visited: Jul 2005.
- Williams, B., Welsh letter-to-sound rules: rewrite rules and two-level rules compared, *Computer Speech & Language*, 8, 261–277, 1994.
- Witten, I. H., A flexible scheme for assigning timing and pitch to synthetic speech, *Language and Speech*, 20, 240–260, 1977.
- Wouters, J., Analysis and synthesis of degree of articulation, Ph.D. thesis, Katholieke Universiteit Leuven, Belgium, 2001.
- Wouters, J., and M. Macon, Effects of prosody on spectral dynamics synthesis, *Journal of the Acoustical Society of America*, 111, 428–438, 2002.
- Wu, C.-H., and J.-H. Chen, Automatic generation of synthesis units and prosody information for Chinese concatenative synthesis, *Speech Communication*, 35, 219–237, 2001.
- Xu, Y., Contextual tonal variations in Mandarin, *Journal of Phonetics*, 25, 61–83, 1997.
- Xu, Y., Consistency of tone-syllable alignment across different syllable structures and speaking rates, *Phonetica*, 55, 179–203, 1998.
- Xu, Y., Effects of tone and focus on the formation and alignment of  $f_0$  contour, *Journal of Phonetics*, 27, 55–105, 1999a.

## BIBLIOGRAPHY

- Xu, Y.,  $f_0$  peak delay: when, where, and why it occurs, [http://home.uchicago.edu/~xuyi/peak\\_delay\\_poster.pdf](http://home.uchicago.edu/~xuyi/peak_delay_poster.pdf), 1999b, visited: May 2004.
- Xu, Y., and X. Sun, Maximum speed of pitch change and how it may relate to speech, *Journal of the Acoustical Society of America*, 111, 1399–1413, 2002.
- Xu, Y., and Q. E. Wang, Pitch target and their realization: evidence from Mandarin Chinese, *Speech Communication*, 33, 319–337, 2001.
- Yang, L.-C., Contextual effects on syllable duration, in *Proc. of 3<sup>rd</sup> ESCA/COCOSDA Workshop on Speech Synthesis*, vol. SSW3-1999, pp. 37–42, ESCA/COCOSDA, Australia, 1998.
- Yao, T. S., G. P. Zhang, and Y. M. Wu, A rule-based Chinese automatic segmentation expert system, *Journal of Chinese Information Processing*, 5, 38–47, 1990.
- Yiourgalis, N., and G. Kokkinakis, A TTS system for Greek language based on concatenation of formant coded segments, *Speech Communication*, 19, 21–38, 1996.
- Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, in *Proceedings of European Conference on Speech Communication and Technology*, vol. 5, pp. 2347–2350, Budapest, 1999.
- Yu, H.-Y., and Y.-H. Oh, Fuzzy expert system for continuous speech recognition, *Expert Systems with Applications*, 9, 81–89, 1995.
- Yu, M.-S., and F.-L. Huang, Disambiguating the senses of non-text symbols for Mandarin TTS systems with a three-layer classifier, *Speech Communication*, 39, 191–229, 2003.
- Yuan, J., C. Shih, and G. P. Kochanski, Comparison of declarative and interrogative intonation in Chinese, in *Proc. of Speech Prosody*, pp. 711–714, Aix-en-Provence, France, 2003.
- Yuan, Y., and M. J. Shaw, Induction of fuzzy decision trees, *Fuzzy Sets & Systems*, 96, 125–139, 1995.
- Zadeh, L. A., A fuzzy-set-theoretical interpretation of linguistic hedges, *J. of Cybernetics*, 2, 4–34, 1972.
- Zera, J., Speech intelligibility measured by adaptive maximum-likelihood procedure, *Speech Communication*, 42, 313–328, 2004.

Part VI  
Appendices

# Appendix A

## Inventory of 230 Yorùbá syllables

This section of the appendix documents the inventory of the Yorùbá syllables disregarding the tones. Each of the 230 syllables is assigned an hexadecimal code. The hexadecimal numbers in the row under CODE column is combined with that above their column number to identifies each syllable. For example the identification code for syllable *a* is 0000 and that for *yu* is 1206.

Table A.1: V and CV type syllable inventory *Total* = 133

CODE	ONSET	RHYME							
		00	01	02	03	04	05	06	
00		a	e	ẹ	i	o	ọ	u	
01	b	ba	be	bẹ	bi	bo	bọ	bu	
02	d	da	de	dẹ	di	do	dọ	du	
03	f	fa	fe	fẹ	fi	fo	fọ	fu	
04	g	ga	ge	gẹ	gi	go	gọ	gu	
05	gb	gba	gbe	gbẹ	gbi	gbo	gbọ	gbu	
06	h	ha	he	hẹ	hi	ho	họ	hu	
07	j	ja	je	jẹ	ji	jo	jọ	ju	
08	k	ka	ke	kẹ	ki	ko	kọ	ku	
09	l	la	le	lẹ	li	lo	lọ	lu	
0A	m	ma	me	mẹ	mi	mo	mọ	mu	
0B	n	na	ne	nẹ	ni	no	nọ	nu	
0C	p	pa	pe	pẹ	pi	po	pọ	pu	
0D	r	ra	re	rẹ	ri	ro	rọ	ru	
0E	s	sa	se	sẹ	si	so	sọ	su	
0F	ṣ	ṣa	ṣe	ṣẹ	ṣi	ṣo	ṣọ	ṣu	
10	t	ta	te	tẹ	ti	to	tọ	tu	
11	w	wa	we	wẹ	wi	wo	wọ	wu	
12	y	ya	ye	yẹ	yi	yo	yọ	yu	

APPENDIX A. INVENTORY OF 230 YORÙBÁ SYLLABLES

Table A.2: Vn and CVn type syllable inventory *Total* = 95

CODE	ONSET	RHYME				
		07	08	09	0A	0B
12		an	en	in	on	un
13	b	ban	b <sub>en</sub>	bin	b <sub>on</sub>	bun
14	d	dan	d <sub>en</sub>	din	d <sub>on</sub>	dun
15	f	fan	f <sub>en</sub>	fin	f <sub>on</sub>	fun
16	g	gan	g <sub>en</sub>	gin	g <sub>on</sub>	gun
17	gb	gban	gb <sub>en</sub>	gbin	gb <sub>on</sub>	gbun
18	h	han	h <sub>en</sub>	hin	h <sub>on</sub>	hun
19	j	jan	j <sub>en</sub>	jin	j <sub>on</sub>	jun
1A	k	kan	k <sub>en</sub>	kin	k <sub>on</sub>	kun
1B	l	lan	l <sub>en</sub>	lin	l <sub>on</sub>	lun
1C	m	man	m <sub>en</sub>	min	m <sub>on</sub>	mun
1D	n	nan	n <sub>en</sub>	nin	n <sub>on</sub>	nun
1E	p	pan	p <sub>en</sub>	pin	p <sub>on</sub>	pun
1F	r	ran	r <sub>en</sub>	rin	r <sub>on</sub>	run
20	s	san	s <sub>en</sub>	sin	s <sub>on</sub>	sun
21	ṣ	ṣan	ṣ <sub>en</sub>	ṣin	ṣ <sub>on</sub>	ṣun
22	t	tan	t <sub>en</sub>	tin	t <sub>on</sub>	tun
23	w	wan	w <sub>en</sub>	win	w <sub>on</sub>	wun
24	y	yan	y <sub>en</sub>	yin	y <sub>on</sub>	yun

Table A.3: N type syllable inventory *Total* = 2

CODE	0C	0D
25	m	n

## Appendix B

### BNF for standard Yorùbá text

#### B.1 Sample Yorùbá text composed for corpus data

##### Ìdàmún Àgbè

Bàbá àgbè ti ta *cocoa* 30 kg ní ₦500 kí ó tó mò pé àjọ *NCB* ti fi owó lé *cocoa*. Ní ìkan bìi dédé agogo 3:00 òsán ni bàbá àgbè délé. Kíá tí àwon alágbèṣe tí ó bá bàbá ṣiṣe ní oko *cocoa* rẹ gbópé óti dé láti ojà ni wón wátò sí enu ònà ilé bàbá àgbè láti gba owó iṣe tí bàbá je wón. Léyin igbà tí bàbá san owó àwon alágbèṣe tán ló tó ri wipé kò sí èrè rárá lórí oun tí òún tà. Díe ni owó tí ó kù fi lé ní ₦102.

Bàbá àgbè wo iṣe ọdún kan tí ó ṣe, àti èrè tí òun je, ni 'bànúnjé bá gba ọkàn rẹ. Nínú ìbànúnjé ni ó fi fi ọrọ náà tó iyálájé létí. Iyálájé mínkanlẹ kí ó tó ṣàlàyé l'ẹkùnréré fún bàbá àgbè. Iyálájé parí ọrọ rẹ pèlú ìkìlò, ó ní; "olè l'àwon tí efi ètò ilú yín le lówó ọ. Ogbónèwé ni wón fi'njálè báyiñ; wón ọ l'òbọn móò! Kí e máa fura, kí esí máa farabalẹ gbó iròyìn inú *radio* lóòrèkóòrè." Léyin àlàyé iyálájé ni bàbá tó mò ìkan tó subúte ọun. Bàbá àgbè mín kanlẹ óní; "háà! àwon àjọ *NBC* yí mà kúkú burú o".

Kò pè kò jìná ni bàbá tún lẹ s'ọjà l'áti lẹ ra dògùn tí yìò fi fín *cocoa*. *Gamalin* 20 sìnì bàbá máa nra tẹlẹ. Díde tó dé *shop* olóògùn *cocoa* kíá l'óyára gbé *gallon* kan tó sí fi ₦65 tó má'un san tẹlẹ lé ọlọjà lówó. Ọlọjà wo bàbá àgbè pèlú ìbínún kí ó tó kégbe l'ọjìjì pé; "sé èyín ṣeṣe dé ilú yí ni? Àbí èyin ọ gbó pé *Gamalin* 20 ti gbówólórí? ₦ 105 ni *Gamalin* 20 kí e yáa gbó! Tí e bá sì pè níbi tí ewà yen, elòmíràn yìò rà mòyín lówó. Kí ele mò, *gallon* 10 ni mò kójáde lóní, eyọ kan ṣo tó kù l'ẹ gbé lówó yen."

Ìbànújé gidi subú lu bàbá àgbè. Ó wòye pé èyí èrè ọun ọ je, èyí ọjà ọun ọ rírá mò. Bàbá mín kanlẹ ó ní; "háà! èyin ọlọjà mà kúkú burú o! Kílódé tí e kó fi jé ká mò kí etó máa f'owó l'ọjà?" Oní *shop* dàhùn óní; "emá báwa wí rárá o! Àwon *government* yín ni kí e lẹ bá. Àbí egbó pé mo ní *factory* *Gamaline* 20 n'ílẹ ni? Oye tí *commissioner* níki ámatáa ni motáa yen. Tí kò bá sì t'èyin l'ọrùn, e lẹ tò s'ára *cocoa* yín!"

Bàbá r'onún jìnlẹ, ó wo oye tí ọun máun ta *cocoa* ní ọdún 1980 àti bí owó dògùn *cocoa* ṣe kéré sí oye tí wón ta *cocoa* nígbà náà sí iye rẹ l'òdun yì. Ó wá ri gbanga pé òwò erú eléèkejì ni ijoba ilú àwon kí àwon àgbè bọ ní ọdún 1992 tí a wàyí. Bàbá wòye pé nígbàti ọun gbónjú, òwe tí àwon àgbà má'n pa nipé, "ayé àgbè, bí ayé à je r'òrun ni". Amó ní bàyí, òwe náà ti yí padà. K'ódà a lè sọpé l'ówó tí a wàyí, "iyà àgbè, ti di iyà àjẹ ròrun."

## B.2 BNF for standard Yorùbá text

<SyDocument>	::= <PARAGRAPH> <PARAGRAPH><PARAGRAPH>
<PARAGRAPH>	::= <Sentence> <Sentence><PARAGRAPH>
<Sentence>	::= <Phrase><DEPUNCT> <Phrase><Sentence>
<Phrase>	::= <word><NONDEPUNCT> <Word><Phrase>
<Word>	::= <Syllable> <Syllable><Word> <Word>
<XWord>	::= <Foreign> <Abbreviation> <Acronym> <Numeral>
	::=  <Symbol>
<Foreign>	::= <Arabic/English/French word> <Proper names>
<Abbreviation>	::= <Letter><.> <Letter><Abbreviation>
<Acronym>	::= <Letter> <Letter><Acronym>
<Numerals>	::= <Time> <Date> <Ordinal><Cardinal><Fraction>
<Time>	::= <Digit><:> <Digit><Time>
<Date>	::= <Year><DELI><Month><DELI><Day>
	::= <Day><DELI><Month><DELI><Year>
	::= <Year><DELI><Day><DELI><Month>
<Ordinal>	::= <Digit> <Digit><Ordinal>
<Cardinal>	::= <Digit> <Digit><Cardinal>
<Fraction>	::= <Digit><.> <Digit><.><Cardinal>
<Syllable>	::= <Base><Tone>
<Tone>	::= <High> <Mid> <Low>
<Base>	::= <Onset> <Rhythm>
<Onset>	::= <Consonant> Null
<Rhythm>	::= <Nuclues><Coda> <Nuclues>
<Nuclues>	::= <Vowel> <Syllabic Nasal> <Nasal>
<Coda>	::= n
<Consonant>	::= {b, d, f, g, gb, h, j, k, l, m, n, p, r, §, S, W, y }
<Nasalised vowel>	::= <sVowel>n
<Vowel>	::= {a, e, e, i, o, o, u }
<sVowel>	::= {a, e, i, o, u }
<Letter>	::= <Consonant> <Vowel>
<Syllabic Nasal>	::= {m, n}
<DEPUNCT>	::= {., ?, !}
<NONDEPUNCT>	::= {,, ;, :, ‘, ’, ‘, ’}
<Symbol>	::= {N, £, \$ }
<Digit>	::= {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
<Deli>	::= {:, -, /}

## Appendix C

# Design of the text markup system for TTS

The text input to a text-to-speech (TTS) synthesis system may come from digital text, including electronic emails, electronic text-books and newspapers. The input may also be in the form of a digital text which specifies the phonological content and structure of the message to be synthesised. In order to generate the speech sound corresponding to the text, a TTS system must extract relevant information from the text. The ease with which this information can be extracted depends on a number of factors, including: the orthography of the target language, the domain described by the text, and whether the content of the text is constrained by some writing rules or standards.

In the SY language orthography, the use of diacritic to represent tone provides additional information about the tone of a syllable. This information is very important in the generation of intonation and prosody for a piece of text. *Bird* (1998) has described the orthography of SY as *shallow*, implying that it is relatively easy to generate pronunciation from the text without the need for complicated syntactic analysis; such as part-of-speech analysis. For example, the pronunciations of the English words *record* (verb) and *record* (noun) differ because of their syntactic classes. This type of situation does not occur in SY.

Despite the closeness of the SY orthography to pronunciation, the amount of information in the text is not enough for producing an accurate prosody for the text. The orthography of SY, and indeed any other language, is designed to provide enough information for a human reader. Paralinguistic aspects of expressions, such as emotion- which endows speech with its naturalness quality- cannot be reduced into writing but are normally guessed by human readers based on context. Other information, such as the structure and style for pronouncing the text are not explicitly represented in the orthography. There is, therefore, the need to augment the content of the text with additional information that will facilitate an unambiguous processing of prosodic data in an input text. In the following subsection we shall discuss the typesetting of SY text that is adopted in this thesis. We will discuss the design of the text markup system for adding additional prosody information into a typeset text.

### C.1 Introduction

A written text can be described as an encoding of a speech utterance using symbols. The text encodes two important information: (i) the content and manner of speech, and (ii) how it should be spoken. The content of the speech, i.e. its written form, is defined by the letters, symbols and numerals in the text. The function of how an entity will be spoken is specified by punctuation marks such as comma (,), full-stop (.), semicolon (;), exclamation mark (!), question mark (?), etc. A space or a hyphen can also be used to indicate entities that are to be pronounced separately or as a discrete unit. For example, a space between two words indicates that each word is a single entity to be pronounced separately; whereas a hyphen used in the same context indicates that the words should be pronounced as if they are a single entity.

The domain of text considered in this work is derived from Yorùbá newspapers and standard



## APPENDIX C. DESIGN OF THE TEXT MARKUP SYSTEM FOR TTS

textbooks. This class of text has two important attributes: (i) a *physical content*, and (ii) a *logical structure*. The physical content is described by tokens which form the component of the text. This includes the following:

- punctuation marks, including spaces;
- lexical items written using the SY orthography;
- numerals;
- symbols and acronyms;
- lexical items written using a foreign orthography (such as English words and proper names).

At a higher level is the grammar that guides the construction and organisation of the symbols and letters into syllables, words, phrases, and sentences.

The logical structure of a text specifies how the text is organised into pronunciation units. The following elements constitute the logical structure of an SY text of reasonable length:

- a title,
- one or more paragraphs,
- sentences,
- phrases,
- words,
- syllables.

The logical structure of an SY document can be described declaratively, and without reference to how a formatting program should realise the desired format.

### C.1.1 Text models

There are three types of text model that can result from the above described SY text structure, namely: (i) *restricted*, (ii) *unrestricted*, and (iii) *markup*. In the restricted text model, specific rules that guide the acceptable contents and structure of the text are explicitly defined. For example, a rule may specify that all numerals or abbreviations in a text must be written in their lexical format. Another rule may specify the type and acceptable location of punctuation marks that must be used in the text. All input text must be written according to the defined rules and hence the input text is forced to conform to a format. The problem with restricted input is that it makes the TTS machine more difficult to use since the input format rules are normally not standardised and are difficult to standardise since there are divergent writing styles and genre of the text, e.g. drama and poem. In addition the text may be written to convey different concepts which will be compromised if the content or style of the text is restricted.

A softcopy of unconstrained text can be represented and stored in a plain text format (e.g. using UNICODE or ASCII). This type of representation is, however, not powerful enough for describing special features of the text that can provide a meaningful input for TTS machines. For example, in plain text, although there are several popular methods, there is no universally agreed convention for delimiting paragraph boundaries.

In the unrestricted text model, there is no restriction on the input text. The text may contain any symbol, letter or number and can represent text from diverse domains such as weather forecast, poems, incantation, etc. The text-preprocessing module of the TTS machine is then required to extract information necessary for synthesising the text. This approach imposes a more complicated text analysis task on the TTS machine. Predicting the prosodic attributes of the text by using automatic or rule based techniques is unlikely to produce an acceptable level of accuracy. *Quené and Kager (1992)* argued that this task can only be accurately done if the automatic process has access to the semantic and pragmatic information about the input text.

In a markup text model, the input text, usually unrestricted, is annotated with information that renders the prosody of the intended utterance transparent. Markup input can easily be built around

a standard markup language, such as the eXtensible Markup Language (XML) making them easy to use by large group of users. Among the information that can be included in an annotated text include intonation phrases, phonetic attributes as well as descriptions of how foreign words and other text anomalies should be pronounced.

Many markup scheme for TTS systems, have been proposed particularly for non-tone languages (*Taylor and Isard, 1997; Ogden et al., 2000*). But such markup schemes are system specific and often use annotation schemes not specifically tailored for tone languages texts.

### C.1.2 Issues in SY typesetting and markup

A very important feature of a fully tone-marked SY text is that the tones and under-dots are adequately indicated, hence tonal information of each syllable can be extracted from the text. This type of text can be typesetted using  $\LaTeX$ . The standard  $\LaTeX$  fonts have all of the components required to compose a Yorùbá text (*Taylor, 2000a*). This is because the  $\LaTeX$  provides markup for indicating diacritic and under-dots associated on letters. This features makes it possible to conveniently generate text with the correct SY orthography. The tones and phones are the two most important ingredients for generating accurate SY pronunciation.

Another feature of SY orthography, which the  $\LaTeX$  system also represents accurately, is the use of unambiguous spaces between words. This information can be used to determine word boundaries. Therefore the typesetting of SY texts in  $\LaTeX$  ensures that accented character format can be used to accurately represent the input to SY TTS.  $\LaTeX$  provides annotation at the lower level of the physical content of text thereby facilitating the incorporation of more accurate utterance production information. In addition,  $\LaTeX$  also facilitates the generation of portable documents, such as PDF and .DVI files, which can be read by humans and efficiently stored and transmitted.

On the word level, however, information about how a word must be pronounced in different contexts, e.g. rate of speech, speaking style, etc., cannot be adequately specified using  $\LaTeX$ . Besides, the logical structure of text which controls the overall prosody of an utterance is better defined at levels high than word (*Quené and Kager, 1992*). Also, the same sequence of words can be read in different manners or styles depending on the context. Phrases, sentences, and paragraphs are not explicitly specified in  $\LaTeX$  except through the use of punctuation marks, such as full-stop, comma, and semi-colon, and other escape sequences like the carriage return. For example, in the case of sentences, full-stop normally used for delimiting declarative statements may be ambiguous in some situations, e.g. in sentences also containing numbers with fractions, for example 12.45.

Predicting the phrase and sentence boundaries is a complicated problem if a  $\LaTeX$  typesetted text is to be processed directly by a TTS system. Moreover, in some speech synthesis tasks, such as text containing dialogue, it may be required to control specific characteristics of the generated speech such as the loudness, rate of speech, age, and gender of the speaker, etc. A TTS system requires phrase and sentence level information to generate the appropriate prosody for a given text. For example, the information on whether a sentence is a question or an exclamation is best defined at sentence and phrase level since they affect utterance units longer than words.

Since  $\LaTeX$  is so effective in describing the physical content of the text, a reasonable approach would be to design another markup system above  $\LaTeX$ , which will describe the logical structure of the text. This will allow us to effectively describe the more abstract higher level prosody of the text.

## C.2 Text-to-Speech markup system

In Chapter 6, we have reviewed the various markup languages and systems commonly used in modern TTS systems. The review reveals two important facts. First, the design of a markup system is greatly influenced by the method selected for the implementation of the TTS high level synthesis module, the features of the target language (e.g. orthography) and the degree of freedom required in the control of prosody in the TTS system. Second, the markup scheme used in most systems are based on the eXtensible Markup Language (XML). This is partly because XML allow users to add additional control commands in a flexible and easy to implement manner. XML provides a more abstract and powerful mean of describing speech prosody at utterance level higher than the syllable

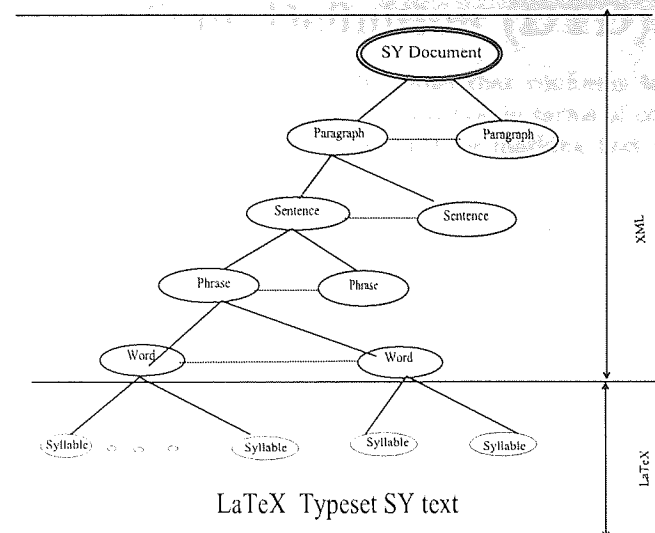


Figure C.1: Example SY Text

(i.e. word, phrase, sentence and paragraph) and facilitates possible publications and data sharing on the Internet.

### C.2.1 Design of XML system

The logical structure of any SY text *document* to be synthesised can be viewed as a tree. The root of the tree is associated with the entire document. The *title* of the document is an optional attribute of the root element. The first sub-tree element is the *paragraph*. A document may contain one or more paragraph elements. Each paragraph is equivalent to a branch from the root document. The second sub-tree element is the *sentence*. A paragraph may contain one or more sentences. The third sub-tree element is the *phrase* and the fifth sub-tree element is the *word*. The phrase element is made up of one or more words and each word element is made up of one or more syllables. Each syllable can be adequately typesetted using  $\LaTeX$  by indicating to diacritic marks and under-dot as required. For example, in syllable  $b\acute{o}$ , the diacritic mark  $\acute{\phantom{o}}$  indicates the tone (i.e. high) and  $b\grave{o}$  is the base. The vowel in the base is the letter  $o$  with an under-dot. Other  $\LaTeX$  specifications for typesetting of SY text are shown in Table C.1.

The XML tree defined above a  $\LaTeX$  document is shown in Figure C.1. The leaf node of the XML tree forms the root of the  $\LaTeX$  part of the tree representation for a piece of text. Element at the same level of the tree are on same level in the linguistic structure of the utterance corresponding to the text. For example, two sentences in the same paragraph of a text share the same branch at the paragraph level. They are also on the same linguistic level in the utterance prosodic hierarchy. The text structure ends with syllable as the leaf elements.

Note that the tree is built from a hierarchical pattern of objects, each of which has a specific attribute which contributes to how the sentence is to be pronounced. We now discuss the design of higher level prosodic markup system using XML.

Table C.1:  $\LaTeX$  annotation for SY diacritic and under dots

Tag	description	Result	Example
$\backslash'$	High tone	$\acute{e}$	Ad $\acute{e}$ (Crown)
$\backslash d$	Under dot	$\grave{e}$	Ql $\grave{e}$ (Fetus)
$\backslash \grave{}$	Low tone	$\grave{e}$	Ql $\grave{e}$ (Lazy)
$\backslash =$ (default)	Mid tone	$\bar{e}$ or (e)	ew $\acute{e}$ (Leaf)

### C.3 Document Type Definition (DTD)

A document type definition (DTD) is a set of declarations that conforms to a particular markup syntax and that describes a class, of “type” of XML documents, in terms of constraints on the logical structure of those document (*Wikipedia*, 2004). In a DTD for marking text for prosody modelling, the structure of a class of documents is described via:

**element definition** tags representing each element required to describe the prosody of SY text, and

**attribute definition** the combination of rules for elements and a list of the information which can be contained within the element. These rules are used to specify the contents of each element.

The beginning of an element is specified by the *start-tag* associated with the tag name and ends with a corresponding *end-tag*. The attribute list for an element contains information which will be included in its tag. An element may have a number of attributes, each of which will have values which can be set when the element is specified. Elements that do not appear in a particular tag may be given a default values

In the design of the tag names for our XML system, we observed that the some annotations  $\LaTeX$  can be confused with standard XML tags. The annotation name specifications in  $\LaTeX$  has the form `\tag_name` while XML tags is of the form `<tag_name>` or `</tag_name>`. But in situation where these names can be confused, we retain the name for  $\LaTeX$  and use the first four upper-case letters of the name for defining the XML tag. Each element is defined by a *start-tag* (e.g. `<document>`) and an end tag (e.g. `</document>`).

There are two important criteria in the design of tag names in the XML our markup system. The first is that computer-literate non-speech-synthesis experts should be able to understand and use the markup. The second is that the markup should be portable across platforms so that a variety of speech synthesis systems should be able to use the additional information provided by the tags *Taylor and Isard* (1997). The general syntax of an XML document is dictated by a set of rules defined by the World Wide Web Consortium (W3C) (*Burnett et al.*, 2002). It consists of a grammar-based set of production rules derived from the Extended Backus-Naur Form (EBNF). In the following, we discuss the design of each tags and illustrate the design using the sample text in Appendix I.

#### C.3.1 The document tag

The `<document>` tag delimits the root element of an SY document. It has an optional *title* attribute which indicates the title of the document. The contents of *title* attribute include an optional text which specifies the title of the document. The `<Document>` tag encloses the text to be spoken. The syntax of the `<Document>` tag is as follows:

```

<document title= "text">
[
<!-- The content of the document will go here -->
[
</document>
```

The plus sign (+) indicates that a document can contain one or more paragraphs. In the case of the text in Appendix I, the `<document>` tag is represented as follows:

```

<document title= "Ìdàmún Àgbẹ" >
{
<!-- The content of the document will go here -->
}
</document>
```

### C.3.2 The paragraph tag

The paragraph tag, <PARA>, defines each paragraph in the document. We use the tag name <PARA> so that it will not be confused with the *paragraph* annotation in L<sup>A</sup>T<sub>E</sub>X. The paragraph contains an optional attribute which indicates the *style* that will be used in uttering the paragraph. The syntax of the <PARA> tag is as follows;

```
<PARA style = "Styletype">
{
sentence+ <!(i.e. The sentences in paragraph) >
}
</PARA>
```

The style attribute accepts five style types:

1. *Dialogue*- a paragraph spoken in the context of a dialogue conversation.
2. *Spontaneous*- a paragraph spoken spontaneously.
3. *Read* - a paragraph spoken as read speech (Default).
4. *Poem* - a paragraph spoken as read Yorùbá poem, e.g. Ewì (common poem), Oríkì (praise song), etc.
5. *Incantation*- a paragraph spoken as SY incantation, e.g. Ofò, Ògèdè, Àsẹ, etc..

The default value for style is *Read*. Using the paragraph tag, the first paragraph in the sample text in Appendix I will be tagged as follows:

```
<document title = Idámún Àgbẹ >
<PARA style= "Read">
<!-- The content of each paragraph will go here -->
</PARA>
</document>
```

### C.3.3 The sentence tag

The sentence tag, <sentence>, delimits an SY sentence and contained in the paragraph element. All sentence elements within the same paragraph have same paragraph attributes. A sentence element contains at least one phrase element. The sentence element has many attributes which specify the prosodic information of a sentence. These attributes are useful in a multi-speaker text, such as a play or a dialogue. It is also required for annotating texts containing more than one reading styles, e.g. a poem. The sentence attributes include the following:

**MODE:** specifies the mode for speaking a sentence. The mode attribute can take one of the following values: *Question, Declaration, Exclamation, Statement*. The default value is *Statement*.

**MOOD:** specifies the mood for the speaking a sentence. The mood attribute can assume one of the following values: *Happy, Sad, Normal*. The default value is *Normal*.

**RATE:** specifies the rate at which a sentence is will be spoken. The attribute has three possible values: *Fast, Normal, Slow*. The default value is *Normal*.

**PAUSE:** signifies the duration of the pause that should be inserted between two words using a linguistic value. This attribute has 3 possible values: *Short, Long, Medium*.

## APPENDIX C. DESIGN OF THE TEXT MARKUP SYSTEM FOR TTS

**LOUD:** signifies the loudness or volume of the sentence as linguistic values. The values possible: *Low, Medium, High*. The default value is medium.

**GENDER:** specifies the type of voice, i.e. male or female, for synthesising the sentence. It has the following values: *Male, Female*. The default value is male.

**AGE:** specifies the approximate age of voice to be synthesised: It has the following values: *Child, Adult, and Old*. The default value is Adult.

The syntax for sentence tagging is therefore:

```
<sentence MODE="Statement" MOOD="Normal" STYLE="Oro"
  RATE="Normal" PAUSE="Medium" LOUD="Medium" >
  <!-- The content of each sentence go here -->
</sentence>
```

All the parameters specified in this syntax are the default values of the attributes. The annotation for the first sentence in our example text is as follows:

```
<sentence MODE="Statement" MOOD="Normal" STYLE="Oro"
  RATE="Normal" PAUSE="Medium" LOUD="Medium" >
  Bàbá àgbè ti ta cocoa 30 kg ní N500 kí ó tó mò pé àjọ NCB ti fi fowó lé cocoa.
</sentence>
```

The last two attributes specifies the kind of speaker's voice to be imitated by the speech synthesiser. If it is not specified an adult male native speaker of SY is assumed. This attribute is only useful for selecting the relevant database in a multi-speaker TTS environment, such as dialogue or story telling. The attribute will guide the TTS engine in selecting the appropriate database.

The sentence tags and attributes discussed above are designed in the manner stated above in order to facilitate the synthesis of text representing a dialogue between two or more people. This situation is very common in plays and newspaper interview texts. The scope of the speaker tag varies. In situation where only one language database is available all specified speaker attributes will be ignored.

### C.3.4 The phrase tag

The phrase tag, `<phrase>`, can be inserted within the text of a sentence to indicate the presence of phrase boundaries. The `<phrase>` effectively breaks a sentence into intonation phrases even when punctuation marks are present. The syntax for the phrase element is as follows:

```
<phrase>
  <!-- The content of each phrase go here -->
</phrase>
```

The `<phrase>` tag for specifying the prosodic phrase for the previous example sentence is as follow:

```
<sentence MODE="Statement" MOOD="Normal" STYLE="Oro"
  RATE="Normal" PAUSE="Medium" LOUD="Medium" >
<phrase> Bàbá àgbè ti ta cocoa 30 kg ní N500 </phrase> <phrase> kí ó tó mò pé
  àjọ NCB ti fi fowó lé cocoa </phrase>.
</sentence>
```

## C.4 Text contents description tags

In a normal SY text, there are many textual anomalies which can occur as part of the content. Some examples of textual anomaly include numerals representing ordinal or cardinal data items, letters representing abbreviations, as well as a groups of letters in foreign -words and proper names, e.g. David. In order to remove the complexities involved in determining the exact pronunciation of these textual anomalies, we defined markup tags for them. These tags are built around the SAYAS tag in W3C (*Burnett et al.*, 2002) and extended to incorporate features specific to SY text. The information provided by this tag will allow the High Level Synthesis module to determine the type of expansion to apply on each of the tagged items during text normalisation process.

### C.4.1 The SAYAS tag

The text content elements are used to markup the content of a text in order to specify how they are to be spoken. For example, ₦500 must be spoken as currency; IADS is to be spoken as an acronym with specific pronunciation, whereas the abbreviation O.A.U. is to be spelt out as English alphabet using Yorùbá accent. The list of tags for content element markup is specified in Table C.2.

Table C.2: Tags for SY text

Token Classification	Description	Tag name
Date	String of numbers formatted as 99-99-99, or 99/99/99	DATE
Time	String of numbers formatted as 99:99, 99/99,	TIME
Currency	String of numbers prefixed by a currency symbol, e.g. ₦, \$, £,	CURRENCY
Lexical	String of letters	LEXICAL
Ordinal digits	String of numbers prefixed by a noun	ORDINAL
Cardinal digit	String of numbers postfixed by a noun	CARDINAL
Loanword	Word with foreign language spelling, e.g. English, French, Arabic, e.t.c.	LOAN
Punctuation	Punctuation marks such as (;) , (:), (.)	PUNCT
Acronym	Group of upper case letters such as FIFA, OAU, USA, e.t.c.	ACRONYM
Special Character	Characters such as *, +, etc.	SPEC
SI unit	SI unit to be expanded into SY accent pronunciation	SUNIT
Phone	Phone number (digit by digit pronunciation)	PHONE
Proper names	Proper names, usually of English origin	PRONAME

Following the W3C format, we use the SAYAS tag with three specific attributes. The SUB attribute is used to specify a substitute text for an abbreviation or acronym. For example, the text *Kóànn* can be substituted for COAN (COMPUTER ASSOCIATION OF NIGERIA). The CLASS attribute is used to specify the class of the pronunciation as stated in Table C.2. When this parameter is not specified, the default is SY abbreviation pronounced with SY accent. Some examples of SAYAS tags is as follows:

```

<SAYAS SUB ="a.b.l."> àti bèè bèè lọ </ SAYAS>
<SAYAS SUB ="i.n.p."> iyen ni pé </ SAYAS>
<SAYAS SUB ="f.w." >fiwé </SAYAS
<SAYAS SUB ="b.a." > bí àpẹẹrẹ </SAYAS>
<SAYAS SUB ="f.a." > fún àpẹẹrẹ </SAYAS
<SAYAS SUB = "w.o.r" > wò ó ore </SAYAS
<SAYAS SUB = "e.n.p." > èyí nipé</SAYAS >

```

### The SUB attribute

The syntax for the SUB attribute is

```
< SAYAS SUB = "text to be substituted" > text </SAYAS>
```

The SUB attribute is particularly useful in defining replacement text for abbreviations and other shorthand text. The substitution strings for some commonly used SY abbreviation are defined as below:

### The CLASS attribute

The syntax for the CLASS attribute is as follows:

```

<SAYAS CLASS = "attribute" > text </SAYAS>
<SAYAS CLASS ="currency" > ₦500</SAYAS>
<SAYAS CLASS ="acronym" > AIDS </SAYAS\>

```

### The ABBRACENT attribute

The third attribute is the abbreviation accent attribute, ABBRACENT. It determines whether an abbreviation will be spelt out as Yorùbá abbreviation using a Yorùbá accent or an English abbreviation (each letter is from the English alphabet) using Yorùbá accent. For example, the abbreviation "a.b.l." (i.e. àti bèè bèè lọ) is a Yorùbá abbreviation and its component alphabet must be pronounced using SY phones. However, O.A.U. (Organisation of African Unit) is an English abbreviation which must be pronounced using SY accent. Below is an example of the usage of the above markup tags:

1. <SAYAS CLASS ="ABBREVIATION" ABBRACENT='English' > O.A.U </SAYAS>

## C.5 Schema for the Document Type Definition (DTD)

```

- <schema xmlns:sytts="http://www.w3.org/2001/XMLSchema">
  xmlns = "http://www.cs.aston.ac.uk/intranet/~odejoboa/sytts"
  targetSpacename=http://www.cs.aston.ac.uk/intranet/~odejoboa/sytts >
  <sytts:Documentation xml:lang="yo" />
  <sytts:element name="document" />
  - <sytts:complexType>
    <sytts:element name="title" minOccurs="1" maxOccurs="unbounded" />
    - <sytts:sequence>
      <sytts:element name="PARA" nype="xsd:complexType" Gender="Male"
        Age="Adult" minOccurs="1" maxOccurs="unbounded" />
    - <sytts:sequence>
      <sytts:element name="sentence" type="xsd:complexType"
        minOccurs="1" maxOccurs="unbounded" mode="xsd:string"

```



APPENDIX C. DESIGN OF THE TEXT MARKUP SYSTEM FOR TTS

```

mood="xsd:string" style="xsd:string" rate="xsd:string" />
<syttps:element name="phrase" type="xsd:complexType"
  minOccurs="1" maxOccurs="unbounded" />
<syttps:element name="word" type="xsd:string" minOccurs="1"
  maxOccurs="unbounded" />
<syttps:element name="lexword" type="xsd:string" minOccurs="1"
  maxOccurs="unbounded" />
<syttps:element name="frword" type="xsd:string" minOccurs="0"
  maxOccurs="unbounded" />
<syttps:element name="numerals" type="xsd:integer" minOccurs="0"
  maxOccurs="unbounded" />
<syttps:element name="currency" type="xsd:float" minOccurs="0"
  maxOccurs="unbounded" />
<syttps:element name="abbreviation" type="xsd:string"
  minOccurs="0" maxOccurs="unbounded" />
<syttps:element name="sui" type="xsd:string" minOccurs="0"
  maxOccurs="unbounded" />
  <syttps:element name="symbol" type="xsd:string" minOccurs="0"
  maxOccurs="unbounded" />
  <syttps:element name="SyllableObject" type="xsd:complexType"
  minOccurs="1" maxOccurs="unbounded" />
</syttps:sequence>
</syttps:sequence>
</syttps:complexType>
</schema>

```

## C.6 Markup of the first paragraph in the sample text

```

<document>
<PARA>
  <sentence><phrase> B\{a}b\{a} \{a}gb\{e} ti ta {cocoa 30 kg n\{i}
    <SAYAS CLASS = 'currency' \sout{N}500> </phrase><phrase>k\{i} \{o} t\{o} m\{d}\{o}
    p\{e} \{a}j\{d}\{o} NCB ti fi fow\{o} l\{e} <SAYAS CLASS = 'LOAN' cocoa>.
  </phrase></sentence>
  <sentence>N\{i} \{n}kan b\{i}i d\{e}d\{e} agogo
    <SAYAS CLASS = 'TIME' 3:00 \d{\{o}}s\{a}n ni b\{a}b\{a} \{a}gb\{e}
    d\{e}l\{e}.
  </sentence>
  <sentence><phrase> K\{i}\{a} t\{i} \{a}won al\{a}gb\{e}\{d}\{s}e
    t\{i} \{o} b\{a} b\{a}b\{a} \{d}\{s}i\{d}\{s}\{d}\{e} n\{i} oko
    <SAYAS CLASS = 'LOAN' cocoa> r\{e} gb\{d}\{o}p\{e} \{o}ti d\{e} l\{a}ti
    \{d}\{o}j\{a}</phrase><phrase> ni w\{d}\{o}n w\{a}t\{o} s\{i} enu \{d}\{o}n\{a}
    il\{e} b\{a}b\{a} \{a}gb\{e} l\{a}ti gba ow\{o} i\{d}\{s}\{e}
    t\{i} b\{a}b\{a} j\{d}\{e} w\{d}\{o}n.
  </phrase></sentence>
  <sentence><phrase>L\{e}y\{i}n \{i}gb\{a} t\{i} b\{a}b\{a}
    \{d}\{s}an ow\{o} \{a}w\{d}\{o}n al\{a}gb\{a}\{d}\{s}e t\{a}n </phrase><phrase>l\{o}
    t\{o} ri w\{i}p\{e} k\{o} s\{i} \{e}r\{e} r\{a}r\{a}
    l\{o}r\{i} oun t\{i} \{o}\{u}n t\{a}.
  </phrase></sentence>
  <sentence>D\{i}\{e} ni ow\{o} t\{i} \{o} k\{u} fi
    l\{e} n\{i} <SAYAS CLASS = 'currency'\sout{N}102>.
  </sentence>
</PARA>
</document>

```

# Appendix D

## Sentences used for modelling

The follow is a list of sentences used in our speech data base for the fundamental frequency and duration modelling. The high tone syllables are tone marked with the acute diacritic mark, e.g. á. The low tone syllables are tone marked with the acute diacritic mark, e.g. à. The mid tone is the default tone, so all mid tone syllables are not tone marked. The sentences comprise forty single phrase sentences and twenty two-phrase sentence. Simple glosses of each of the sentences are also provided. The glosses are intended to provide the closes possible English interpretation of the sentences.

### D.1 Single phrase sentences

1. Bàbá àgbè ti ta kòkó.  
The farmer has sold cacao.
2. Ó mò pé èmi kò.  
He knows that it is not me.
3. Wọn ti fowó lé ojà.  
They have increased the price of commodity.
4. Iṣé ló wáwá.  
He came look for job.
5. Ìyálójà ló mò.  
The head of the market women (female) knows.
6. Adé orí oba.  
The crown in the king's head.
7. Ìwà ló yàtò.  
It is behaviour that differs.
8. Enun ọ̀nà ló gbé sí.  
He placed it by the door.
9. Dídé tódé ló wá síbí.  
He came here as soon as he arrived.
10. Wọn ti mowó wá.  
They have brought the money.
11. Olùkó ti dé.  
The teachers has come.
12. Ó jìn sí kòtò.  
He fell into a ditch.
13. Ati parí.  
We have finished.

APPENDIX D. SENTENCES USED FOR MODELLING

14. Àgbà òsikà.  
Grumpy old man (woman).
15. Omọ wéwé ni wón.  
They are kids.
16. Èlérù ti dé.  
The owner of the baggage has come.
17. Ìlú tí kò l'óba.  
A town without a king.
18. Bótiwí lóri.  
It is as he has said.
19. Ó tún sáré wá.  
He came back running again.
20. Ònà yí jìn.  
The road is far.
21. Ìwé ni mò íkọ.  
I am writing.
22. Ó ti mbò.  
He is on the way.
23. Ìjọba ti sòfin tuntun.  
The government has promulgated a new law.
24. Ìyá ni wúrà.  
The mother is precious.
25. Ìwò aṣẹ ja.  
The fisher's nest.
26. Isu aṣẹ ja.  
The fisher's yam.
27. Oúnjẹ ti délẹ.  
The food is ready.
28. Kò tètè jí.  
He woke up late.
29. Ìwà ọ̀dèlẹ.  
A behaviour of betrayer.
30. Dide díró.  
Stand up.
31. Aràrá gùnkè odò.  
The dwarf climbed up the stream.
32. Oko Àdìgún lati s̄is̄é.  
We worked at Adigun's farm.
33. Kététété òkè ọya ló gùn re lé.  
He rides on the Donkey from over the niger.
34. Láì f̄ọ̀rọ̀ gùn.  
Without prolonging the issue.
35. Ó mún afára gùn.  
He climbed the bridge.
36. Lójú ẹ̀sẹ̀ ni ọpa ịsẹ̀ tó òsẹ̀ tì.  
He stopped what he was doing immediately.

## APPENDIX D. SENTENCES USED FOR MODELLING

37. Ìyálájé minkanlè.  
The head (female) of the market women took a deep breadth.
38. Iṣé ilé náà ti parí.  
The building work is completed.
39. Kòsì òtító nínún ọ̀rọ̀ náà.  
There is no truth in the word (i.e. word spoken by someone).
40. Efi àṣiṣe rẹ̀ hànán.  
Show him his mistake.

### D.2 Two phrase sentences

1. Bàbá àgbẹ̀ ti ta kòkó, kótó mò pé kòkó ti gbówó lórí.  
The farmer has sold his cocoa, before known that the price of cocoa has increased.
2. Ọ̀dọ̀mi ló kọ̀kọ̀ sáwá, kí ó tó rìn padà sí ilú Ìbàdàn.  
He ran to my place, before running back to *Ibadan* town.
3. Opẹ̀ kí ó tó dé sí ilé, nítorí ọ̀nà tó jín ló ti rìn wá.  
He came home late, because he walked from a far place.
4. Ó mò pé ẹ̀mi kọ̀ ló gbé àga ìyá àgbà, sórí igi ẹ̀mi.  
He knows that I am not the one, that put the old woman's chair on the *emi* tree.
5. Àlàdé wá, ọ̀sì tún lọ.  
*Alade* came, and left as well.
6. Ọ̀nà yí jìn, a kò lerímí.  
This road is far, we cannot walk it.
7. Elẹ̀rù ti dé, ẹ̀jẹ̀ ká gbẹ.  
The owner of the baggage has come, let us carry it.
8. Láì fọ̀rọ̀ gùn, ó mún afára gùn.  
Without prolonging the issue, he climbed the bridge.
9. Olùkọ̀ wá, wón sì kọ̀ wa.  
The teacher came, and he taught us.
10. Ìyá ti dé, ẹ̀jẹ̀ ká jẹun.  
Mother has come, let us eat.
11. Ibi tí a lọ, latí m̀bò.  
We are coming from where we went.
12. Léyín igbìyànjú ogún ọ̀dún, iṣé ilé náà ti parí.  
After twenty years of toil, the building work has completed.
13. Gégé bí akọ̀wé ti wí, kò sí idí tí afi nílátì lọ.  
As the secretary has said, there is no reason for us to go.
14. Orí igi ọ̀pẹ̀ ni Àlàdé gùn, nígbà tí ó rí ẹ̀jò.  
*Alade* climbed the palm tree, when he saw a snake.
15. Ọ̀gbẹ̀ni Gbàdà, wá fi orúkọ̀ sílẹ̀ fún káàdì idánimò.  
Mr. *Gbada*, come and register for identification card.
16. Ìròyìn tó tẹ̀wá lówọ̀ ni pé, kòsì òtító nínún ọ̀rọ̀ náà.  
The news reaching us is that, there is no truth in the account.
17. Ọ̀sìṣẹ̀ ìlera fi kún ọ̀rọ̀ rẹ̀ pé, kò sí òògùn fún àisàn ifòyà.  
The health worker further said that, there is no cure for fear.

*APPENDIX D. SENTENCES USED FOR MODELLING*

18. Fún àpẹẹrẹ, kò sí ìlú tí kò ní ìjọba.  
For example, there is no town without a king.
19. Èfi àṣiṣẹ̀ rẹ̀ hànán, kí ẹ̀ sì tọ̀ sọ̀nà.  
Show him his mistake, and reprimand him.
20. Sìbẹ̀ sìbẹ̀, ó lọ láìi gbàṣe.  
Even then, he left without taking approval.

# Appendix E

## Labelled speech data

This appendix contains example annotated speech files used for this research. All the speech files are hand labelled. Figures E.1, E.2 and E.3 show the syllable files annotations together with the spectrogram of each syllable. The next three figures, i.e. , Figures E.4, E.5 and E.6 show one-phrase sentence annotation files. The last two figures, i.e. Figure E.8 and E.7 shows two-phrase sentence annotation files.

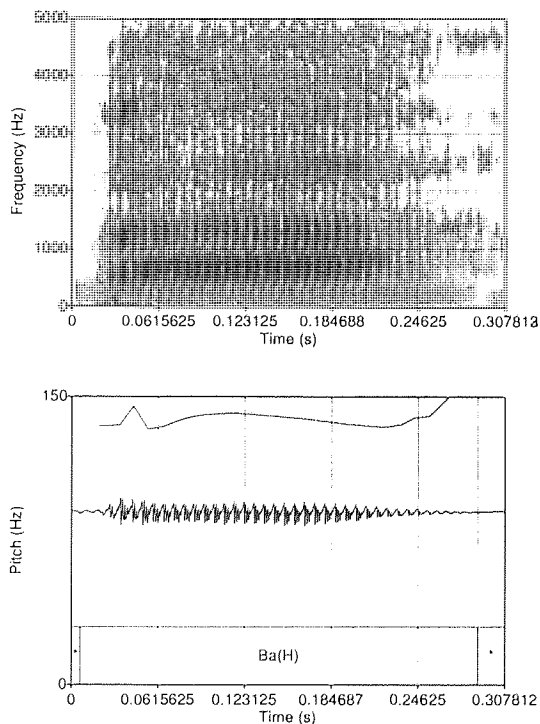


Figure E.1: SY syllable Bá (get to)

APPENDIX E. LABELLED SPEECH DATA

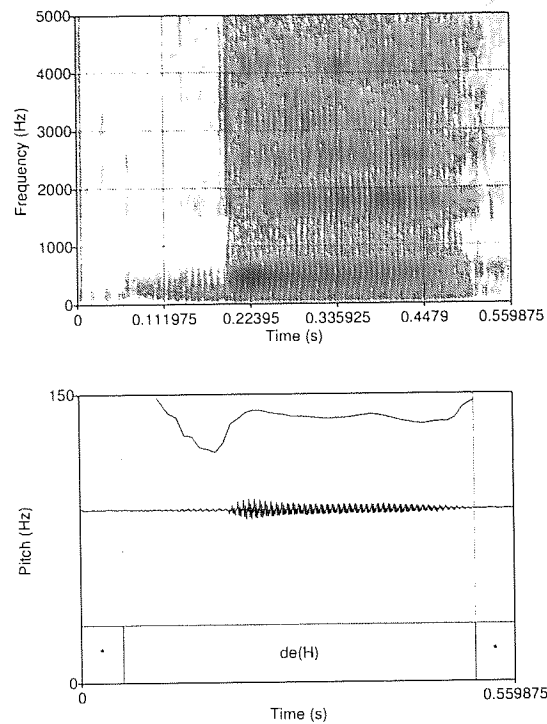


Figure E.2: SY syllable Dé (cover)

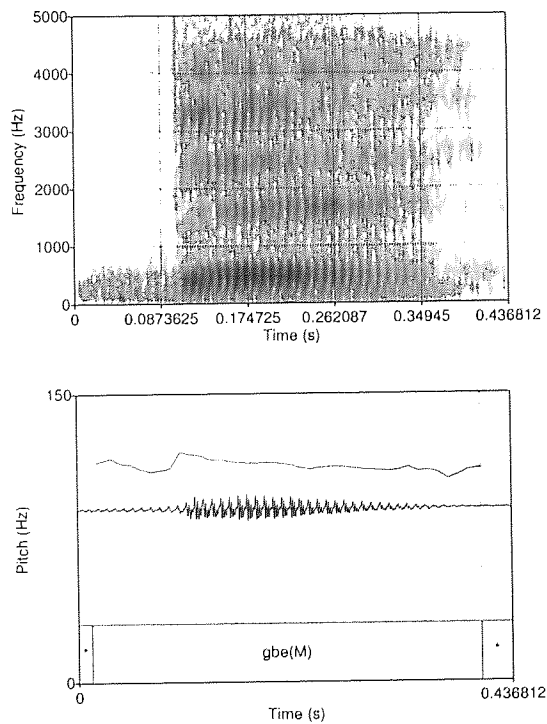


Figure E.3: SY syllable Gbe (to dig)



APPENDIX E. LABELLED SPEECH DATA

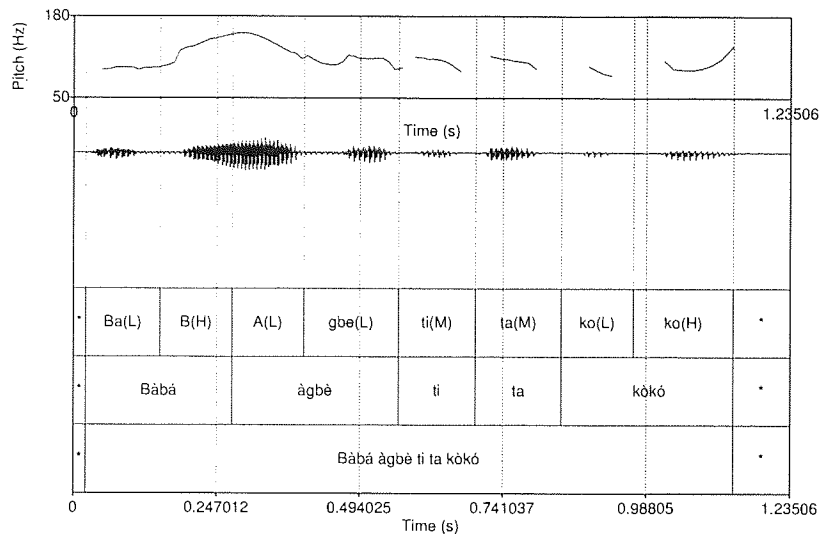


Figure E.4: Annotate file for the SY sentence "Bábá àgbè tí ta kòkó."

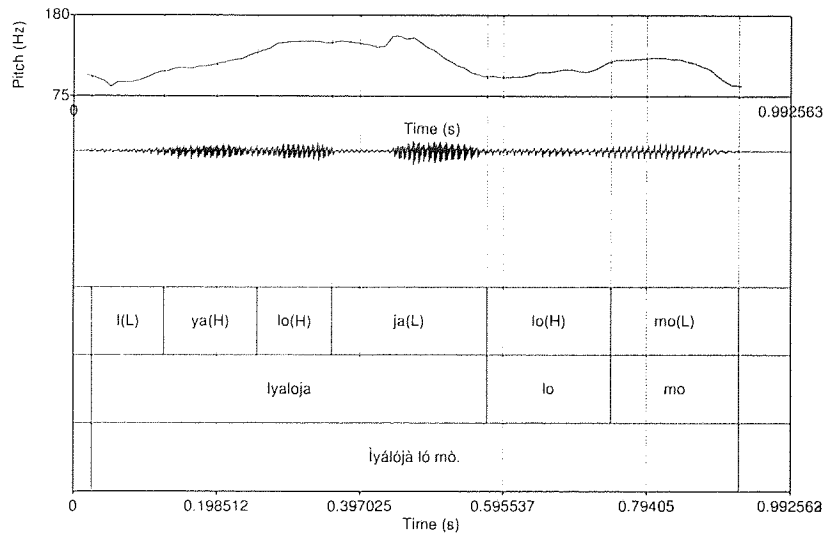


Figure E.5: Annotate file for the SY sentence "Iyálójà ló mò."

APPENDIX E. LABELLED SPEECH DATA

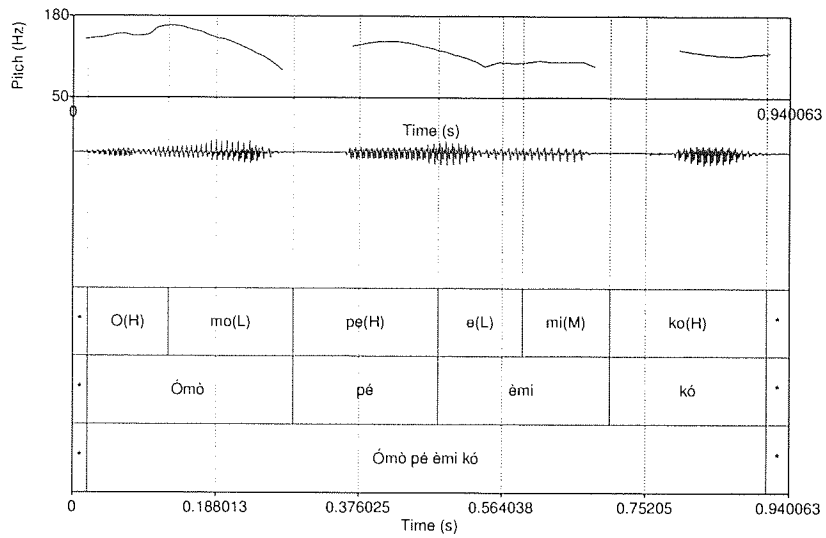


Figure E.6: Annotate file for the SY sentence "Ó mò pé èmi kó."

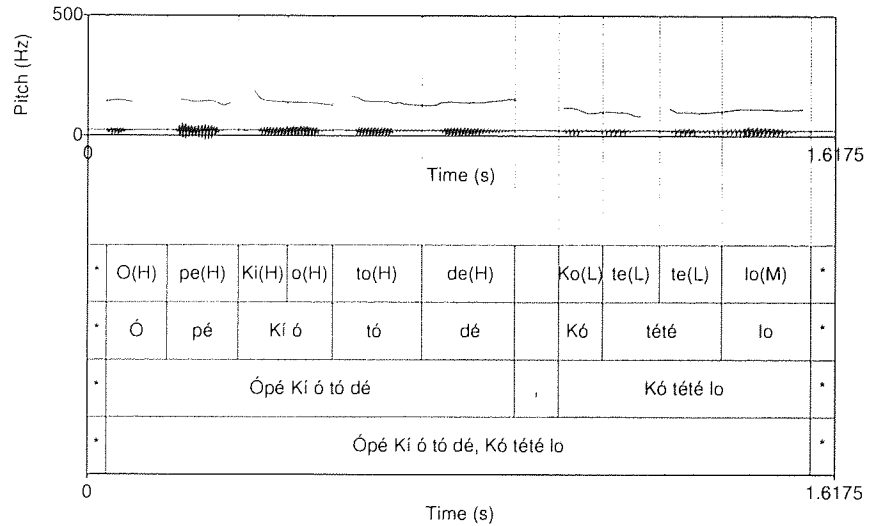


Figure E.7: Annotate file for the SY sentence "Ópé kí ó tó dé, kò tètè lo."

APPENDIX E. LABELLED SPEECH DATA

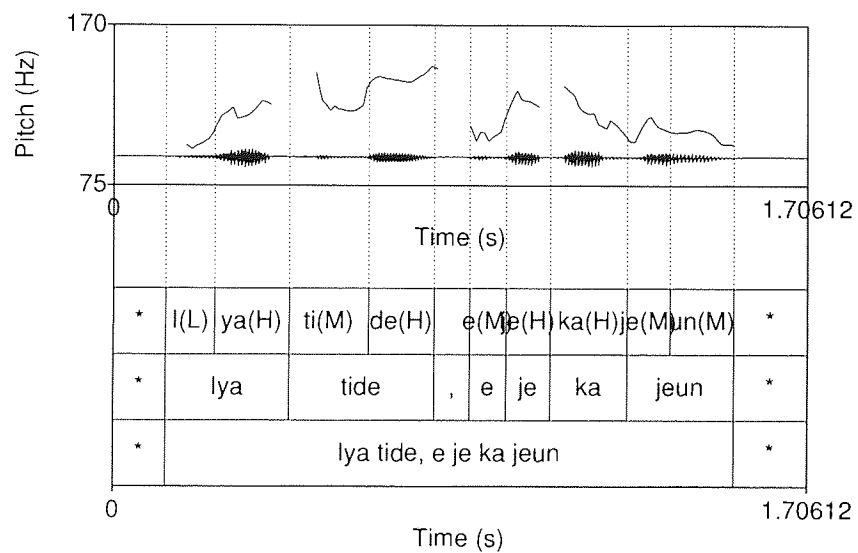


Figure E.8: Annotate file for the SY sentence “*Iyá tidè, ejé ká jeun*”.

# Appendix F

## Program listings

The listing of some the programs used in this research are as follow:

### F.1 Program for extracting data from annotation speech files

The following C code is used to extract data from the annotated speech database and format them for prosody modelling. Depending on the parameter supplied to the program, it will generate the duration and  $f_0$  data for each syllable in the input file. the programme is also used to generate the input file for the Stem-ML program.

```
#include <stdio.h>
#include <stdlib.h>
#include <ctype.h>
#include <string.h>

void getPTone(char , char *, int, char);
void writeHead();

void readInput( FILE *);

FILE *fininput, *foutput;

char Exc1[30];

float number1[30], number2[30], wscale[30], pos[30];

int main(int argc, char *argv[]) {
    char c1[5]= " ";
    char lastc1 = ' ';
    int k, r, pindex=0, Spos=0, K2;
    float number1[30], number2[30], wscale[30], pos[30];
    char ac_name = ' ';
    char ltype[4], cname[6], strenght[10];
    strcpy(ltype, "tag");

    if (argc <2){
        fprintf(stderr, "Missing processing file \n");
        exit(1);}
    if((fininput = fopen(argv[1], "r") ) == NULL){
        fprintf(stderr, "Cannot open input file");
        exit(1);}

        fprintf(stderr, "Opening output file \n");
        foutput = fopen("Outputfile", "a");
        if(argv[2] != NULL) {
```

## APPENDIX F. PROGRAM LISTINGS

```

        fprintf(stderr, "Writing header \n");
        writeHead();}
readInput(fininput);
fclose(foutput);
fclose(fininput);
return 0;
}

void getPTone(char Dac_name, char *Dstrength, int Dpos, char
Dlastc1) {

    if (Dac_name == 'H' )strcpy(Dstrength, "(p14)");
    if (Dac_name == 'M' )strcpy(Dstrength, "(p15)");
    if (Dac_name == 'L' )strcpy(Dstrength, "(p16)");
    if (Dlastc1 == '*' ) strcat(Dstrength,"+(p17)");
    if (Dlastc1 == '.' ) strcat(Dstrength,"+(p19)");
    if (Dlastc1 == ',' ) strcat(Dstrength,"+(p18)");
}

void writeHead(){
    fprintf(foutput, "add=(0.5); ltype=settings;\n");
    fprintf(foutput, "ltype=settings; smooth=(0.1*math.exp((p0)));\n");
    fprintf(foutput, "base=(p1)*100.0; ltype=settings;\n");
    fprintf(foutput, "ltype=settings; max =350;\n");
    fprintf(foutput, "ltype=settings; min =20;\n");
    fprintf(foutput, "ltype=settings; pdroop =0;\n");
    fprintf(foutput, "adroop =(0); ltype=settings;\n");
    fprintf(foutput, "jitter=0.0; ltype=settings;\n");
    fprintf(foutput, "jittercut=1.0; ltype=settings;\n");
    fprintf(foutput, "ltype=settings; range=100;\n");
    fprintf(foutput, "atype=(p2); ctrshift=(p3); ltype=aclass; name=H;
shape=[(-0.500, p4),(0.500,p5)] ; stype=(0); type=(p6); wscale=(p7);\n");
    fprintf(foutput, "atype=(p2); ctrshift=(p3); ltype=aclass; name=M;
shape=[(-0.500, p8),(0.500,p9)] ; stype=(0); type=(p10);wscale=(p7);\n");
    fprintf(foutput, "atype=(p2); ctrshift=(p3); ltype=aclass; name=L;
shape=[(-0.500,p11),(0.500,p12)]; stype=(0); type=(p13); wscale=(p7);\n");
    fprintf(foutput, "ltype=fom; fom=(p6*0.001 )\n");
    fprintf(foutput, "ltype=fom; fom=(p10*0.001)\n");
    fprintf(foutput, "ltype=fom; fom=(p13*0.001)\n;\n");
}

void readInput(FILE *FilePT){
    char Tc1[5] = " "; int j =0; int k, n, r,l; float num1, num2;
    fprintf(stderr, "Reading input file \n");
    while(!feof(FilePT)) {
        fscanf(FilePT, "%f %f %s", &number1[j], &number2[j], Exc1[j]);
        printf("%f \n",number1[j]);
        printf("%f\n",number2[j]);
        l = strlen(Tc1); Tc1[l] = '\0';
        printf("%i %s\n", l, Exc1[j]);
        k = strlen(Tc1);
        for(r=0; r<k; ++r){
            if (Tc1[r] == '(') Exc1[j] = Tc1[r+1];
            if(Tc1[k-1]== ',' ) Exc1[j] = Tc1[k-1];
            ++j;} }
    fprintf(stderr, "Printing input file \n");
    for(n=0; n <j; ++n){
        fprintf(stderr, "%f %f %s", number1[n], number2[n], Exc1[n]);
    }
}

```

## F.2 MatLab programs

### F.2.1 Stylistation interpolation program

The following program is used to generate the interpolation into the  $f_0$  values of syllable during the stylistation experiments. The programme also calculates the root mean square error of each plot.

```
H =[129.980163574 124.321533203 121.813079834 122.074432373 ...
    123.565002441 134.247558594 134.00869751 ...
    134.975341797 144.831161499 123.181495667 134.382659912
    136.798019409 ...
    136.593841553 138.462402344 ...
    138.750656128 138.776733398 137.632766724 138.740493774
    138.794509888 ...
    138.79083252 138.403884888 ...
    139.486816406 137.828216553 136.022903442 135.529708862
    132.665298462 ...
    132.504196167 138.849685669 138.787246704 136.189971924];

M= [109.738960266 112.350662231 115.24962616 133.506622314 ...
    115.227027893 114.961486816 114.996704102 ...
    115.138542175 115.888816833 115.966300964 116.497322083 116.467674255 ...
    116.592193604 115.367195129 ...
    115.015159607 114.621025085 113.008583069 114.148178101 110.640907288 ...
    108.355995178 104.601257324 109.353164673];

L =[124.90411377 113.839149475 114.417953491 114.377365112 ...
    113.686386108 111.161354065 110.331420898 ...
    111.280250549 120.963867188 113.085090637 108.480064392 ...
    99.3661575317 96.0092697144 93.2647399902 ...
    90.2358474731 87.3196105957 85.5085144043 82.2133026123 ...
    82.4408416748 77.7474822998 77.5981674194 ...
    76.630279541 74.5484085083];

yh = length(H);
xh = 1: yh;

ym = length(M);
xm = 1: ym;

yl = length(L);
xl = 1: yl;

subplot(2, 2, 4);

p1 = polyfit(xh, H, 3);
p2 = polyfit(xm, M, 3);
p3 = polyfit(xl, L, 3);

apH = polyval(p1, xh);
apM = polyval(p2, xm);
apL = polyval(p3, xl);
erh1 = 0.0;
erh2 = 0.0;
erh3 = 0.0;

for j = 1: yh
    erh1 = erh1 + (H(j) - apH(j)) * (H(j) - apH(j));
end

for j = 1: ym
    erh2 = erh2 + (M(j) - apM(j)) * (M(j) - apM(j));
end
for j = 1: yl
    erh3 = erh3 + (L(j) - apL(j)) * (L(j) - apL(j));
end
erh1 = sqrt(erh1)/yh erh2 = sqrt(erh2)/ym erh3 = sqrt(erh3)/yl
```

## APPENDIX F. PROGRAM LISTINGS

```

subplot(2,2,1); plot(xh,apH, '- ',xm, apM, '- ',xl, apL,'.-' ); title
('Cubic interpolation into values F0 for Yoruba Syllable /Gba/');
xlabel ('Frame # '); ylabel ('F0 Values in Hz' );

p1 = polyfit(xh, H, 4);

p2 = polyfit(xm, M, 4);

p3 = polyfit(xl, L, 4);

apH = polyval(p1, xh);
apM = polyval(p2, xm);
apL = polyval(p3, xl);
erh1 = 0.0;
erh2 = 0.0;
erh3 = 0.0;

for j = 1: yh
    erh1 = erh1 + (H(j) - apH(j)) * (H(j) - apH(j));
end

for j = 1: ym
    erh2 = erh2 + (M(j) - apM(j)) * (M(j) - apM(j));
end

for j = 1: yl
    erh3 = erh3 + (L(j) - apL(j)) * (L(j) - apL(j));
end

erh1 = sqrt(erh1)/yh;

erh2 = sqrt(erh2)/ym;

erh3 = sqrt(erh3)/yl;

subplot(2,2,2); plot(xh,apH, '- ',xm, apM, '- ',xl, apL,'.-' ); title
('Forth Order interpolation into values F0 for Yoruba Syllable
/Gba/'); xlabel ('Frame # '); ylabel ('F0 Values in Hz' );

p1 = polyfit(xh, H, 5);

p2 = polyfit(xm, M, 5);

p3 = polyfit(xl, L, 5);

apH = polyval(p1, xh);
apM = polyval(p2, xm);
apL = polyval(p3, xl);
erh1 = 0.0; erh2 = 0.0; erh3 = 0.0;

for j = 1: yh
    erh1 = erh1 + (H(j) - apH(j)) * (H(j) - apH(j));
end

for j = 1: ym
    erh2 = erh2 + (M(j) - apM(j)) * (M(j) - apM(j));
end

for j = 1: yl
    erh3 = erh3 + (L(j) - apL(j)) * (L(j) - apL(j));
end

erh1 = sqrt(erh1)/yh

erh2 = sqrt(erh2)/ym;

```

## APPENDIX F. PROGRAM LISTINGS

```

erh3 = sqrt(erh3)/y1;

subplot(2,2,3); plot(xh,apH, '- ',xm, apM, '- ',xl, apL,'.-' ); title
('Fifth order interpolation into values F0 for Yoruba Syllable
/Gba/'); xlabel ('Frame # '); ylabel ('F0 Values in Hz' );

p1 = polyfit(xh, H, 6);
p2 = polyfit(xm, M, 6);
p3 = polyfit(xl, L, 6);

apH = polyval(p1, xh);
apM = polyval(p2, xm);
apL = polyval(p3, xl);
erh1 = 0.0;
erh2 = 0.0;
erh3 = 0.0;

for j = 1: yh
    erh1 = erh1 + (H(j) - apH(j)) * (H(j) - apH(j));
end

for j = 1: ym
    erh2 = erh2 + (M(j) - apM(j)) * (M(j) - apM(j));
end

for j = 1: yl
    erh3 = erh3 + (L(j) - apL(j)) * (L(j) - apL(j));
end

erh1 = sqrt(erh1)/yh;
erh2 = sqrt(erh2)/ym;
erh3 = sqrt(erh3)/yl;

subplot(2,2,4);

plot(xh,apH, '- ',xm, apM, '- ',xl, apL,'.-' ); title ('Sixth Order
interpolation into values F0 for Yoruba Syllable /Gba/'); xlabel
('Frame # '); ylabel ('F0 Values in Hz' );

```

The following program implements the intonation models discussed in *Láníran and Clements* (2003) as that presented by *Shih* (2000). It also generate a plot which compare the two intonation patterns with the one we developed in this research.

```

Tonesequce =[107 105 110 107 107 107 107 98 95 107 107 99 107 107 89
107 107 100 107 107];
TonesequcT =['M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M'
'M' 'M' 'M' 'M' 'M' 'M'];
k = length(Tonesequce);
LaniraP(1) = Tonesequce(1);
ShishP(1) =
Tonesequce(1);

T2p(1) = Tonesequce(1);
NaturalT2p(1) = Tonesequce(1);

xk = 1:k;
Alpha = 0.40;
Beta = 0.92;
Step = 0.09;

Lp = Tonesequce(1);
Cp = Tonesequce(1);

```



## APPENDIX F. PROGRAM LISTINGS

```

p = Tonesequce(1);

d =0.65;
RefH= 100;
RefM = 89;
RefL = 65;

for j = 2:k
    Lp(j) = 0.6*(Lp(j-1)- RefM) + RefM;
    Cp(j) = 0.9*(Cp(j-1)) + 0.76*(Tonesequce(1) - 0.9*Tonesequce(1));
    T2p(j) = Lp(j) + j*0.12;
    NaturalT2p(j) = T2p(j)*0.98 + j*0.01;
end

plot(xk, Lp, '-.', xk, Cp, ':', xk, T2p, '*-', xk,NaturalT2p,'o-');

Tonesequce =[78 78 66 71 81 86 71 69 75 67 70 80 63 73 64 80
69 72 70 71]; TonesequcT =['L' 'L' 'L' 'L' 'L' 'L' 'L' 'L' 'L' 'L'
'L' 'L' 'L' 'L' 'L' 'L' 'L' 'L' 'L'];

Tp =[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]; Lp =[0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0]; Cp =[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0];
%TonesequcV
Tv =[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];%[ 0 0 0 0 0 0 0];
TonesequcN =[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];%[ 0 0 0 0 0 0 0];
TonesequcNv=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];%[ 0 0 0 0 0 0 0];
TonesequcV=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];

Alpha = 0.40; Beta = 0.92; Step = 0.06;

d =0.6;
RefH = 100;
RefM = 89;
RefL = 69;

TonesequcN(1) = Tonesequce(1);
TonesequcNv(1) = TonesequcV(1);

LastH = 0.0;
LastM = 0.0;
LastL = 0.0;

if TonesequcT(1) == 'H'
    LastH = Tonesequce(1);
end

if TonesequcT(1) == 'M'
    LastM = Tonesequce(1);
end

if TonesequcT(1) == 'L'
    LastL = Tonesequce(1);
end

k = length(Tonesequce);
xk = 1:k;

Tp(1) = Tonesequce(1);
Lp(1) = Tonesequce(1);

Cp(1) = Tonesequce(1);
Tv(1) = TonesequcV(1);

for j = 2: k
    if TonesequcT(j) == 'H'
        if LastH == 0.0
            LastH = Tonesequce(j);
        end
    end
end

```

## APPENDIX F. PROGRAM LISTINGS

```
end
TonesequcT(j);
Tp(j) = Alpha*(((Tonesequc(j)/LastH))*(LastH - RefH)) +
Beta*(1 - (LastH/Tp(j-1))) + RefH

Tp(j) = Alpha*(((Tonesequc(j)/LastH))*(LastH - RefH)) +
Beta*(((Tp(j-1)/LastH)) + RefH;
LastH = Tp(j);
Tv(j) = TonesequcV(j) *(LastH/Tonesequc(j));
end

if TonesequcT(j) == 'M'
if LastM == 0.0
LastM = Tonesequc(j);
end
Tp(j) = Alpha*(((Tonesequc(j)/LastM))*(LastM - RefM)) +
Beta*(((Tp(j-1)/LastM)) + RefM;
LastM =Tp(j);
Tv(j) = TonesequcV(j) *(LastM/Tonesequc(j));
end
if TonesequcT(j) == 'L'
if LastL == 0.0
LastL = Tonesequc(j);
end
TonesequcT(j);
Tp(j) = Alpha*(((Tonesequc(j))/(LastL))*(LastL - RefL)) +
Beta*(((Tp(j-1)/LastL)) + RefL;
LastL =Tp(j);
Tv(j) = TonesequcV(j) *(LastL/Tonesequc(j));
end

Alpha = Alpha - Step;

Beta = Beta + Step;

TonesequcN(j) = Tp(j) + Tp(j)*j*0.011;

TonesequcNv(j) = Tv(j) + Tv(j)*j*0.012;
end

TonesequcN(j) = 62;

TonesequcNv(j) = 50;

%subplot(2,2,4);
subplot(1,2,2) subplot(1,2,1);

plot(xk, Tp, '-.', xk, TonesequcN, '*-'); title ('Synthesise and
natural F0 peak for utterance contour'); xlabel ('Frame # '); ylabel
('F0 Values in Hz ');
```

### F.3 Praat program listings

The algorithms for the stylisation, duration modelling and speech synthesis experiments are coded in the *Praat* script. The program that implements the stylisation algorithms was used to experiment with various  $f_0$  stylisation functions. This *Praat* script goes through sound and TextGrid files in a directory and opens each pair of Sound and TextGrid. The Sound file is converted into a Manipulation Object and the maximum and minimum  $f_0$  value of each labelled interval is calculated.

The  $f_0$  curve is then replaced by linear, quadratic, cubic, 4<sup>th</sup> order, 5<sup>th</sup> order, or 6<sup>th</sup> order interpolation polynomial depending on the option input into the program. The sound resulting from the re-synthesised Manipulation Object is stored for perceptual experiment. This program consists of about 350 lines of *Praat* codes. The interface to that program is shown in Figure F.1.

The *Praat* duration script calculates the total duration of the intervals in a selected tier which have a regular label (i.e. not those labelled with '\*'). The location, on the duration tier, of the

## APPENDIX F. PROGRAM LISTINGS

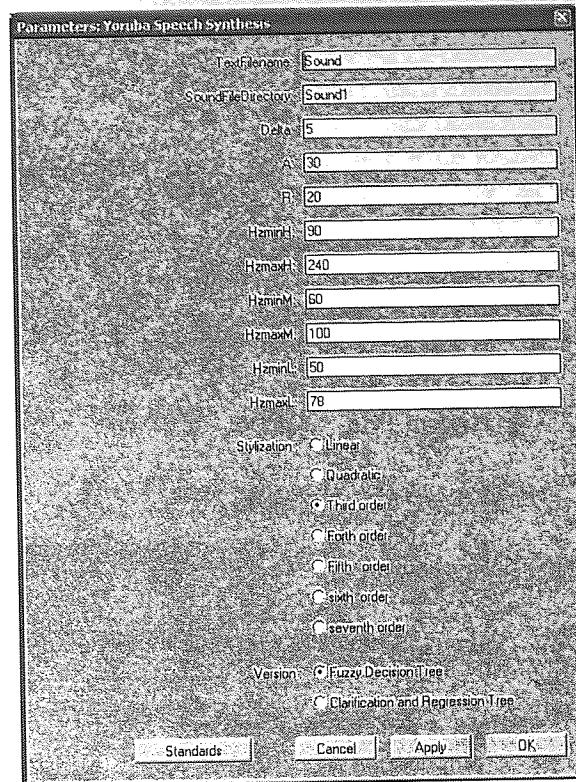


Figure F.1: A screen capture of the interface to the *Praat* programs.

minimum and maximum  $f_0$  value is also replaced by those computed the stylisation program. The Fuzzy Decision Tree or Classification and Regression Tree option can be selected to experiment with the desired duration modelling technique. The duration tier for each labelled item is replaced by the computed duration. The sound resulting from the re-synthesised Manipulation Object is stored for perceptual experiment. This program consists of about 280 lines of *Praat* codes.

The *Praat* script for implementing the prosody synthesis algorithm uses function from the stylisation and duration script to calculate the parameter of the speech sound to be synthesised. In order to generate the synthesised sound, a manipulation object is created and the the duration and  $f_0$  dimensions computed by the duration and intonation modelling modules used to replaced those on the Manipulation Object. The Manipulation Object is then stored for perceptual and other evaluation. This program consists of 390 lines of *Praat* codes. Most of the program discussed above are modified from *Mietta Lenne(2002)*.