DEVELOPMENTS IN SPEECH SYNTHESIS

Mark Tatham

Department of Language and Linguistics, University of Essex, UK

Katherine Morton

Formerly University of Essex, UK

John Wiley & Sons, Ltd

DEVELOPMENTS IN SPEECH SYNTHESIS

DEVELOPMENTS IN SPEECH SYNTHESIS

Mark Tatham

Department of Language and Linguistics, University of Essex, UK

Katherine Morton

Formerly University of Essex, UK

John Wiley & Sons, Ltd

Copyright © 2005

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-85538-X (HB)

Typeset in 10/12pt Times by Graphicraft, Limited, Hong Kong, China.
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Acknowledgements **xiii**

Introduction **1**
How Good is Synthetic Speech? 1
Improvements Beyond Intelligibility 1
Continuous Adaptation 2
Data Structure Characterisation 3
Shared Input Properties 4
Intelligibility: Some Beliefs and Some Myths 5
Naturalness 7
Variability 8
The Introduction of Style 10
Expressive Content 11
Final Introductory Remarks 13

Part I Current Work **15**

1 High-Level and Low-Level Synthesis **17**
1.1 Differentiating Between Low-Level and High-Level Synthesis 17
1.2 Two Types of Text 17
1.3 The *Context* of High-Level Synthesis 18
1.4 Textual Rendering 20

2 Low-Level Synthesisers: Current Status **23**
2.1 The Range of Low-Level Synthesisers Available 23
2.1.1 Articulatory Synthesis 23
2.1.2 Formant Synthesis 24
2.1.3 Concatenative Synthesis 28
Units for Concatenative Synthesis 28
Pepresentation of Speech in the Database 31
Unit Selection Systems: the Data-Driven Approach 32
Unit Joining 33
Cost Evaluation in Unit Selection Systems 35
Prosody and Concatenative Systems 35
Prosody Implementation in Unit Concatenation Systems 36
2.1.4 Hybrid System Approaches to Speech Synthesis 37

3 Text-To-Speech **39**
3.1 Methods 39
3.2 The Syntactic Parse 39

4 Different Low-Level Synthesisers: What Can Be Expected? **43**
4.1 The Competing Types 43
4.2 The Theoretical Limits 45
4.3 Upcoming Approaches 45

5 Low-Level Synthesis Potential **47**
5.1 The Input to Low-Level Synthesis 47
5.2 Text Marking 48
5.2.1 Unmarked Text 48
5.2.2 Marked Text: the Basics 48
5.2.3 Waveforms and Segment Boundaries 50
5.2.4 Marking Boundaries on Waveforms: the Alignment Problem 51
5.2.5 Labelling the Database: Segments 54
5.2.6 Labelling the Database: Endpointing and Alignment 55

Part II A New Direction for Speech Synthesis **57**

6 A View of Naturalness **59**
6.1 The Naturalness Concept 59
6.2 Switchable Databases for Concatenative Synthesis 60
6.3 Prosodic Modifications 61

7 Physical Parameters and Abstract Information Channels **63**
7.1 Limitations in the Theory and Scope of Speech Synthesis 63
7.1.1 Distinguishing Between Physical and Cognitive Processes 64
7.1.2 Relationship Between Physical and Cognitive Objects 65
7.1.3 Implications 65
7.2 Intonation Contours from the Original Database 65
7.3 Boundaries in Intonation 67

8 Variability and System Integrity **69**
8.1 Accent Variation 69
8.2 Voicing 72
8.3 The Festival System 74
8.4 Syllable Duration 75
8.5 Changes of Approach in Speech Synthesis 76

9 Automatic Speech Recognition **79**
9.1 Advantages of the Statistical Approach 80
9.2 Disadvantages of the Statistical Approach 81
9.3 Unit Selection Synthesis Compared with Automatic Speech Recognition 81

Part III High-Level Control **83**

10 The Need for High-Level Control **85**
10.1 What is High-Level Control? 85

10.2 Generalisation in Linguistics 86
10.3 Units in the Signal 89
10.4 Achievements of a Separate High-Level Control 90
10.5 Advantages of Identifying High-Level Control 90

11 The Input to High-Level Control 93
11.1 Segmental Linguistic Input 93
11.2 The Underlying Linguistics Model 94
11.3 Prosody 96
11.4 Expression 98

12 Problems for Automatic Text Markup 99
12.1 The Markup and the Data 100
12.2 Generality on the Static Plane 101
12.3 Variability in the Database–or Not 102
12.4 Multiple Databases and Perception 105
12.5 Selecting Within a Marked Database 105

Part IV Areas for Improvement 109

13 Filling Gaps 111
13.1 General Prosody 111
13.2 Prosody: Expression 112
13.3 The Segmental Level: Accents and Register 113
13.4 Improvements to be Expected from Filling the Gaps 115

14 Using Different Units 119
14.1 Trade-Offs Between Units 119
14.2 Linguistically Motivated Units 119
14.3 A-Linguistic Units 121
14.4 Concatenation 123
14.5 Improved Naturalness Using Large Units 123

15 Waveform Concatenation Systems: Naturalness and Large Databases 127
15.1 The Beginnings of Useful Automated Markup Systems 129
15.2 How Much Detail in the Markup? 129
15.3 Prosodic Markup and Segmental Consequences 132
15.3.1 Method 1: Prosody Normalisation 132
15.3.2 Method 2: Prosody Extraction 133
15.4 Summary of Database Markup and Content 135

16 Unit Selection Systems 137
16.1 The Supporting Theory for Synthesis 137
16.2 Terms 138
16.3 The Database Paradigm and the Limits of Synthesis 139
16.4 Variability in the Database 139
16.5 Types of Database 140
16.6 Database Size and Searchability at Low-Level 142
16.6.1 Database Size 142
16.6.2 Database Searchability 144

Part V Markup **145**

17 VoiceXML **147**
17.1 Introduction 147
17.2 VoiceXML and XML 148
17.3 VoiceXML: Functionality 148
17.4 Principal VoiceXML Elements 149
17.5 Tapping the Autonomy of the Attached Synthesis System 151

18 Speech Synthesis Markup Language (SSML) **153**
18.1 Introduction 153
18.2 Original W3C Design Criteria for SSML 153
Consistency 153
Interoperability 154
Generality 154
Internationalisation 154
Generation and Readability 155
Implementability 155
18.3 Extensibility 155
18.4 Processing the SSML Document 155
18.4.1 XML Parse 156
18.4.2 Structure Analysis 156
18.4.3 Text Normalisation 157
18.4.4 Text-To-Phoneme Conversion 157
18.4.5 Prosody Analysis 159
18.4.6 Waveform Production 160
18.5 Main SSML Elements and Their Attributes 160
18.5.1 Document Structure, Text Processing and Pronunciation 160
18.5.2 Prosody and Style 161
18.5.3 Other Elements 162
18.5.4 Comment 162

19 SABLE **165**

20 The Need for Prosodic Markup **167**
20.1 What is Prosody? 167
20.2 Incorporating Prosodic Markup 167
20.3 How Markup Works 168
20.4 Distinguishing Layout from Content 168
20.5 Uses of Markup 169
20.6 Basic Control of Prosody 170
20.7 Intrinsic and Extrinsic Structure and Salience 172
20.8 Automatic Markup to Enhance Orthography: Interoperability with the Synthesiser 174
20.9 Hierarchical Application of Markup 175
20.10 Markup and Perception 176
20.11 Markup: the Way Ahead? 177
20.12 Mark What and How? 179
20.12.1 Automatic Annotation of Databases for Limited Domain Systems 180
20.12.2 Database Markup with the Minimum of Phonology 180
20.13 Abstract Versus Physical Prosody 182

Part VI Strengthening the High-Level Model **183**

21 Speech **185**
21.1 Introductory Note 185
21.2 Speech Production 186
21.3 Relevance to Acoustics 186
21.4 Summary 187
21.5 Information for Synthesis: Limitations 187

22 Basic Concepts **189**
22.1 How does Speaking Occur? 189
22.2 Underlying Basic Disciplines: Contributions from Linguistics 191
22.2.1 Linguistic Information and Speech 191
22.2.2 Specialist Use of the Terms 'Phonology' and 'Phonetics' 192
22.2.3 Rendering the Plan 193
22.2.4 Types of Model Underlying Speech Synthesis 194
The Static Model 194
The Dynamic Model 194

23 Underlying Basic Disciplines: Expression Studies **197**
23.1 Biology and Cognitive Psychology 197
23.2 Modelling Biological and Cognitive Events 198
23.3 Basic Assumptions in Our Proposed Approach 198
23.4 Biological Events 198
23.5 Cognitive Events 201
23.6 Indexing Expression in XML 203
23.7 Summary 204

24 Labelling Expressive/Emotive Content **207**
24.1 Data Collection 208
24.2 Sources of Variability 209
24.3 Summary 210

25 The Proposed Model **213**
25.1 Organisation of the Model 213
25.2 The Two Stages of the Model 214
25.3 Conditions and Restrictions on XML 214
25.4 Summary 215

26 Types of Model **217**
26.1 Category Models 217
26.2 Process Models 218

Part VII Expanded Static and Dynamic Modelling **219**

27 The Underlying Linguistics System **221**
27.1 Dynamic Planes 221
27.2 Computational Dynamic Phonology for Synthesis 222
27.3 Computational Dynamic Phonetics for Synthesis 223
27.4 Adding *How*, *What* and Notions of Time 224

27.5 Static Planes 224
27.6 Computational Static Phonology for Synthesis 225
27.7 The Term *Process* in Linguistics 226
27.8 Computational Static Phonetics for Synthesis 228
27.9 Supervision 230
27.10 Time Constraints 230
27.11 Summary of the Phonological and Phonetic Models 231

28 Planes for Synthesis 233

Part VIII The Prosodic Framework, Coding and Intonation 235

29 The Phonological Prosodic Framework 237
29.1 Characterising the Phonological and Phonetic Planes 239

30 Sample Code 245

31 XML Coding 249
31.1 Adding Detail 250
31.2 Timing and Fundamental Frequency Control on the Dynamic Plane 256
31.3 The Underlying Markup 257
31.3.1 Syllables and Stress 258
31.3.2 Durations 260
31.4 Intrinsic Durations 261
31.5 Rendering Intonation as a Fundamental Frequency Contour 262
1: Assign Basic *f0 Values to All S and F Syllables in the Sentence: the Assigned Value is for the Entire Syllable* 263
2: Assign f0 for all U Syllables; Adjust Basic *Values* 263
3: Remove Monotony 264
4: For Sentences with RESET, where a RESET Point is a Clause or Phrase Boundary 264

32 Prosody: General 265
32.1 The Analysis of Prosody 266
32.2 The Principles of Some Current Models of Intonation Used in Synthesis 268
32.2.1 The Hirst and Di Cristo Model (Including INTSINT) 268
32.2.2 Taylor's Tilt Model 269
32.2.3 The ToBI (Tones and Break Indices) Model 269
32.2.4 The Basis of Intonation Modelling 270
32.2.5 Details of the ToBI Model 271
32.2.6 The INTSINT (International Transcription System for Intonation) Model 273
32.2.7 The Tatham and Morton Intonation Model 274
Units in T&M Intonation 274

33 Phonological and Phonetic Models of Intonation 277
33.1 Phonological Models 277
33.2 Phonetic Models 277
33.3 Naturalness 278
33.4 Intonation Modelling: Levels of Representation 281

Part IX Approaches to Natural-Sounding Synthesis **283**

34 The General Approach **285**
34.1 Parameterisation 285
34.2 Proposal for a Model to Support Synthesis 286
34.3 Segments and Prosodics: Hierarchical Ordering 287
34.4 A Sample Wrapping in XML 288
34.5 A Prosodic Wrapper for XML 289
34.6 The Phonological Prosodic Framework 290

35 The Expression Wrapper in XML **291**
35.1 Expression Wrapping the *Entire* Utterance 292
35.2 Sourcing for Synthesis 293
35.3 Attributes Versus Elements 294
35.4 Variation of Attribute Sources 296
35.5 Sample Cognitive and Biological Components 297
35.5.1 Parameters of Expression 298
35.5.2 Blends 298
35.5.3 Identifying and Characterising Differences in Expression 298
35.5.4 A Grammar of Expressions 299

36 Advantages of XML in Wrapping **301**
36.1 Constraints Imposed by the XML Descriptive System 303
36.2 Variability 303

37 Considerations in Characterising Expression/Emotion **305**
37.1 Suggested Characterisation of Features of Expressive/Emotive Content 305
37.1.1 Categories 305
37.1.2 Choices in Dialogue Design 307
37.2 Extent of Underlying Expressive Modelling 308
37.3 Pragmatics 309

38 Summary **313**
38.1 Speaking 313
38.2 Mutability 315

Part X Concluding Overview **317**

Shared Characteristics Between Database and Output: the Integrity of the Synthesised Utterance 319
Concept-To-Speech 321
Text-To-Speech Synthesis: the Basic Overall Concept 322
Prosody in Text-To-Speech Systems 323
Optimising the Acoustic Signal for Perception 325
Conclusion 326

References **329**

Author Index **335**

Index **337**

Acknowledgements

We should like to acknowledge British Telecom Research Labs, IBM-UK Research, the UK Department of Trade and Industry, and the UK Engineering and Physical Sciences Research Council (formerly the Science Research Council) for early support for our work on the integration of cognitive and physical approaches in the speech synthesis environment. Thanks in general are due to colleagues in the area of computer modelling of speech production and perception, and in particular to Eric Lewis (Department of Computer Science, University of Bristol).

Introduction

How Good is Synthetic Speech?

In this book we aim to introduce and discuss some current developments in speech synthesis, particularly at the higher level, which focus on some specific issues. We shall see how these issues have arisen and look at possible ways in which they might be dealt with. One of our objectives will be to suggest that a more unified approach to synthesis than we have at the present time may result in overall improvement to synthesis systems.

In the early days of speech synthesis research the obvious focus of attention was *intelligibility*–whether or not the synthesiser's output could be understood by a human listener (Keller 1994; Holmes and Holmes 2001). Various methods of evaluation were developed which often involved comparison between different systems. Interestingly, intelligibility was almost always taken to mean *segmental intelligibility*–that is, whether or not the speech segments which make up words were sufficiently well rendered to enable those words to be correctly recognised. Usually tests for intelligibility were not performed on systems engaged in dialogue with humans–the test environment involved listeners evaluating a synthesiser just speaking to them with no interaction in the form of dialogue. The point here is that intelligibility varies with context, and a dialogue simulation would today be a much more appropriate test environment for intelligibility.

It is essential for synthesisers to move away from the basic requirements of minimally converting text-to-speech–see Dutoit (1997) for a comprehensive overview–to systems which place more emphasis on naturalness of speech production. This will mean that the earlier synthesis model will necessarily become inadequate as the focus shifts from the reading task *per se* to the quality of the synthetic voice.

Improvements Beyond Intelligibility

Although for several years synthetic speech has been fully intelligible from a segmental perspective, there are areas of naturalness which still await satisfactory implementation (Keller 2002). One area that has been identified is *expressive content*. When a human being speaks there is no fixed prosodic rendering for particular utterances. There are many ways of speaking the same sentence, and these are dependent on the various features of expression. It is important to stress that, whatever the source of expressive content in speech, it is an extremely changeable parameter. A speaker's expression varies within a few words, not just from complete utterance to complete utterance. With the present state of

Developments in Speech Synthesis Mark Tatham and Katherine Morton
© 2005 John Wiley & Sons, Ltd. ISBN: 0-470-85538-X

the art it is unlikely that a speech synthesiser will reflect *any* expression adequately, let alone one that is varying.

But to sound completely natural, speech synthesisers will sooner or later have to be able to reflect this most natural aspect of human speech in a way which convinces listeners that they could well be listening to real speech. This is one of the last frontiers of speech synthesis–and is so because it constitutes a near intractable problem.

There is no general agreement on what naturalness actually is, let alone on how to model it. But there are important leads in current research that are worth picking up and consolidating to see if we can come up with a way forward which will show promise of improved naturalness in the future. The work detailed in this book constitutes a *hypothesis*, a proposal for pushing speech synthesis forward on the naturalness front. It is not claimed in any sense that we are presenting the answer to the problem.

Many researchers agree that the major remaining obstacle to fully acceptable synthetic speech is that it continues to be insufficiently natural. Progress at the segmental level, which involves the perceptually acceptable rendering of individual segments and how they conjoin, has been very successful, but *prosody* is the focus of concern at the moment: the rendering of suprasegmental phenomena–elements that span multiple segments–is less than satisfactory and appears to be the primary source of perceptual unease. Prosody itself however is complex and might be thought of as characterising not just the basic prosody associated with rendering utterances for their plain meaning, but also the prosody associated with rendering the expressive content of speech. Prosody performs multiple functions– and it is this that needs particular attention at the moment. In this book one concern will be to address the issue of correct, or appropriate prosody in speech–not just the basic prosody but especially the prosody associated with expression.

Does synthetic speech improve on natural speech? According to some writers, for example Black (2002), there is a chance that some of the properties of speech synthesis can in fact be turned to advantage in some situations. For example, speech synthesisers can speak faster, if necessary, than human beings. This might be useful sometimes, though if the speech is faster than human speech it might be perceived or taken to be of lower quality. Philosophically this is an important point. We have now the means to convey information using something which is akin to human speech, but which could actually be considered to be an *improvement* on human speech. For the moment, though, this looks suspiciously like an explanation after the fact–turning a bug into a hidden feature! But this wouldn't be the first time in the history of human endeavour when we have had to admit that it possible to improve on what human beings are capable of.

Continuous Adaptation

The voice output of current synthesis systems does not automatically adapt to particular changes that occur during the course of a dialogue with a human being. For example, a synthetic utterance which begins with fast speech, ends with fast speech; and one which begins sounding firm does not move to a gentler style as the dialogue unfolds. Yet changes of this kind as a person speaks are a major property of naturalness in speech.

To simulate these changes for adequate synthesis we need a data structure characterisation sufficiently detailed to be able to handle dynamic changes of style or expression during the course of an utterance. We also need the means to introduce *marking* into the utterance specification which will reflect the style changes and provide the trigger for the appropriate procedures in the synthetic rendering.

The attributes of an utterance we are focussing on here are those which are rendered by the prosodic structure of the utterance. Prosody at its simplest implements the rhythm, stress and intonational patterns of canonical utterances. But in addition the parameters of prosody are used to render expressive content. These parameters are often characterised in a way which does not enable many of the subtleties of their use in human speech to be carried over to synthesis. For example, rate of delivery can vary considerably during the course of an utterance–a stretch of speech which might be characterised in linguistic terms as, say, a phrase or a sentence. Rate of delivery is a physical prosodic parameter which is used to render different styles that are characterised at an abstract level. For example, angry speech may be delivered at a higher than normal rate, bored speech at a lower than normal rate.

Take as an example the following utterance:

The word I actually used was *apostrophe*, though I admit it's a bit unusual.

In the orthographic representation of the word *apostrophe*, italicisation has been used to highlight it to indicate its infrequent use. In speech the word might

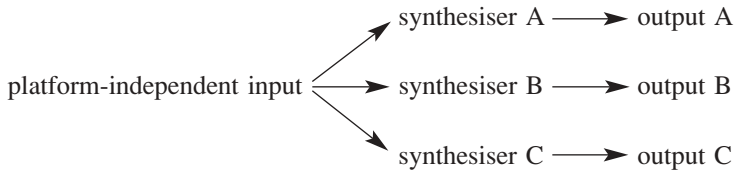
- be preceded and followed by a pause
- be spoken at a rate lower than the surrounding words
- have increased overall amplitude, and so on.

These attributes of the acoustic signal combine to throw spoken highlighting onto the word, a highlighting which says: *this is an unusual word you may not be familiar with*. In addition, uttering the word slowly will usually mean that phenomena associated with fast delivery (increased coarticulation, deliberate vowel reduction etc.) may not be present as expected. To a great extent it is the violation of the listener's expectations–dictated largely by the way the sentence has begun in terms of its prosodic delivery–which signals that they must increase their attention level here. What a speaker expects is itself a variable. By this we mean that there is a norm or baseline expectation for these parameters, and this in itself may be relative. The main point to emphasise is the idea of *departure from expectation*–whatever the nature or derivation of the expectation. In a sense the speaker *plays on* the listener's expectations, a concept which is a far cry from the usual way of thinking about speakers. We shall be returning frequently to the interplay between speaker and listener.

Data Structure Characterisation

Different synthesis systems handle both segmental and prosodic phenomena in different ways. We focus mainly on prosody here, but the same arguments hold for segmental phenomena. There is a good case for characterising the objects to be rendered in synthetic speech identically no matter what the special properties of any one synthesis system. Platform independence enables comparison and evaluation beyond the idiosyncrasies of each system.

Identical input enables comparison of the differing outputs, knowing that any differences detected have been introduced during the rendering process. For example:



Note that outputs A, B and C can be compared with each other with respect to the common input they have rendered. Differences between A, B and C are therefore down to the individual characteristics of synthesisers A, B and C respectively.

The evaluation paradigm works only if the synthesisers under scrutiny are compliant with the characteristics of the input. Each may need to be fronted by a conversion process, and the introduction of this stage of processing is itself, of course, able to introduce errors in the output. But provided care is taken to ensure a minimum of error in the way the synthesiser systems enable the conversion the paradigm should be sound.

Applications in the field may need to access different synthesisers. In such a case a platform-independent high-level representation of utterances will go a long way to ensuring a minimum of disparity between outputs sourced from different systems. This kind of situation could easily occur, for example, with call centres switching through a hierarchy of options which may well involve recruiting subsystems that are physically distant from the initiating controller. The human enquirer will gain from the accruing continuity of output. However, common input, as we have seen, does not guarantee identity of output, but it does minimise discontinuities to which human users are sensitive. Indeed, since there are circumstances in which different synthesis systems are to be preferred–no single one is a universal winner–it helps a lot with comparison and evaluation if the material presented to all systems is identical.

Shared Input Properties

The properties in the input that are common are those which are quite independent of any subsequent rendering in the synthesis process. In general these properties are regarded as linguistic in nature, and in the linguistics model they precede phonetic rendering for the most part. By and large it is phonetic rendering which the low-level synthesis system is simulating. However, synthesis systems which incorporate high-level processing, such as text-to-speech systems, include phonological and other processing. Design of a platform-independent way of representing input to systems which incorporate some high-level processing is much more difficult than systems which involve only phonetic rendering. There are two possible polarised solutions, and several in between.

1 Remove from the text-to-speech system all processing more appropriately brought to a markup of the input text.
2 Introduce a platform-specific intermediate stage which removes only the least successful parts of the text-to-speech system and/or identifies any ‘missing’ processes.

The first solution implies standardising on high-level processes like text normalisation and orthography-to-phoneme conversion. This may not be a bad thing provided designers of text-to-speech systems were not compelled to drop their own processing in these areas if this were both compatible and at least as successful with respect to the final output. The problem really arises when we drop below this simple textual processing level and move into the linguistics processing proper–the phonology and in particular the prosody. We shall see later that whatever approach is adopted it becomes essential to identify what is to be drawn from an input markup and what is to be supplied in the text-to-speech system itself. The main way of avoiding confusion will be to have a common set of level identifiers to be used as markers indicating where in individual systems this or that process (to be included or rejected) occurs.

It will turn out to be important in the way we model human and synthetic speech production to distinguish between the linguistic properties of speech and the two main ways of rendering those properties: by human beings or by synthesisers. Within each of these two types there are different subtypes, and once again it is helpful if the input to all types is characterised in the same way. Along with linguistics in general, we claim a universality for the way in which all phenomena associated with speech production which *can* be characterised within linguistics (and perhaps some areas of psychology if we consider the perception of speech also) are to be described. What this means simply is that much is to be gained from adopting a universal framework for characterising important aspects of speech perception and production by both human beings and computers.

Intelligibility: Some Beliefs and Some Myths

A fairly common hypothesis among synthesis researchers is that the intelligibility of synthetic speech declines dramatically under conditions that are less than ideal. It is certainly true that when listening conditions are adverse synthetic speech appears to do less well than human speech as far as listeners are concerned–prompting the notion that human speech has more critical detail than synthetic speech. It follows from this that it can be hypothesised that adding the missing detail to synthetic speech will improve its intelligibility under adverse or more realistic listening conditions.

It is not self-evident, though, that increased detail is what is needed. For example it may well be that some systematic variation in human speech is not actually perceived (or used in the perception process) and/or it may be the case that some *non*-systematic detail is perceived, in the sense that if it is missing the result is the assertion that the speech is not natural. What constitutes naturalness is not entirely clear to anyone yet, if we go into this amount of detail in trying to understand what it means for speech to be intelligible to the point of being natural. Hence it becomes easy to equate naturalness with increased intelligibility and assign both to improved detail in the acoustic signal. If correct, then we go full circle on the observation that somehow or other human speech holds on to its intelligibility under adverse conditions where synthetic speech does not–even though both may be judged equally intelligible in the laboratory. Assertions of this kind are not helpful in telling us exactly what that detail of human speech might be; they simply inform us that human speech is perceptually more robust than synthetic speech, and that this is perhaps surprising if the starting point–the perception in the laboratory environment–is apparently equal. In our model (see Part V Chapter 20 and Part IX) we would hypothesise that the

perceptual assignment process involves inputs that are not wholly those which on the face of it are responsible for intelligibility.

It is not difficult to imagine that formant synthesis may be producing a soundwave which is less than complete. The parameters for formant synthesis were selected in the early days–for example in the Holmes (1983) model–based on their obviousness in the acoustic signal and on their hypothesised relevance to perception. Thus parameters like formant peak frequency, formant amplitude and formant bandwidth were seen to be important, and were duly incorporated. Later systems–for example the Klatt (1980) synthesiser–built on this model to include parameters which would deliver more of the acoustic detail while attempting to maintain the versatility of formant or parametric synthesis. Fundamentally, it is quite true that, however carefully formant synthesis models the acoustic production of speech, the resultant signal is inevitably lacking in coherence and integrity. A speech signal with 100% integrity would require an infinite number of parameters to simulate it. The robust correlation between vocal tract behaviour and the detail of the acoustic signal is what makes natural speech acoustically coherent: its acoustic fine detail reflects vocal tract behaviour and identifies the signal as coming from a single talker. Indeed we could go further: the correlation is not just robust, it is probably *absolute*. What this correlation does not do on its own, however, is guarantee *phonetic* coherence, since vocal tract behaviour has a nonlinear relationship with phonetics and includes unpredictable cognitively sourced elements (Morton 1986; Tatham 1986a).

One or two researchers have taken a state-of-the-art parametric device–such as the Klatt synthesiser–and made the theoretical assumption that the coherence of its *output* can be improved by working on the internal integrity of its *input* (Stevens and Bickley 1991). HLSyn (Stevens 2002) is one such attempt. The proponents of HLSyn propose a level of representation which is intermediate between what is generally called high-level synthesis (corresponding to phonological, prosodic and pragmatic planning of utterances in linguistics) and low-level synthesis–the actual parametric device which creates the soundwave. They confuse the issue somewhat by calling HLSyn 'high-level synthesis', which is an idiosyncratic use of the term *high*. We shall see later that HLSyn and other comparable approaches (Werner and Haggard 1969; Tatham 1970a) do indeed introduce added coherence by linking acoustic detail *via* a shared higher level of representation – in this case an articulatory level (see also Mermelstein 1973). We would argue that for the moment it has not been shown that a similar level of coherence can be introduced by simply organising the acoustic parameters into an integrated structure.

Our own philosophy works roughly along these lines too: we are concerned with the integrity of the high-level parts of synthesis (rather than the intermediary levels which concern the HLSyn researchers). The principal example of this is our approach to prosody and expression–insisting that all utterance plans be wrapped in tightly focussed prosodic containers which ultimately control the rendering of temporal and spectral features of the output signal whether this is derived from a formant model or a concatenative waveform model.

But, certainly in our own experience, there are also similar though less severe problems with concatenated waveform synthesis, even in those systems which attempt to optimise unit length. This leads us to believe that although, of course, a certain minimum level of acoustic detail is necessary in all synthetic speech, the robustness issue is not down solely to a failure to replicate the greater spectral detail of human speech. What is left, of course, is prosodic detail and temporally governed variation of spectral detail. We are referring here to subtlety in fundamental frequency contours, and variations in intensity and rhythm for

the prosodic detail *per se*; and also to the way spectral detail (for example, the variation in coarticulatory effects and the way they span much more than just the immediately adjacent segments) is governed by features like rate variation. These features are very complex when considering prosody in general, but particularly complex when considering prosody as conveyor of expression.

Naturalness

We shall be referring often in this book to natural sounding speech, and we share with many others an awareness of the vagueness of this idea. The perceived feeling of naturalness about speech is clearly based on a complex of features which it is difficult to enumerate. The reason for this is that listeners are unable to tell us precisely what contributes to naturalness. Several researchers have tried to introduce a metric for naturalness which goes beyond the simple marking of a scale, and introduces the notion of a parametric characterisation of what people feel as listeners. While not new of course in perceptual studies, such a method does go a long way toward enabling comparison between different systems by establishing the basis for a rough evaluation metric.

Take for example the naturalness scoring introduced by Sluijter *et al*. (1998). The approach is technically parametric and enumerates eleven parameters which listeners are asked to consider on five-point scales. They refer to these as a measure of acceptability, but acceptability and naturalness begin to converge in this type of approach–because of the idea that what is acceptable is also a prerequisite for what is natural. Sluijter *et al*.'s parameters can be readily glossed, adapted and extended:

1 *General quality*. What general impression does the speech create? In many studies this is the overall concept of naturalness and very often the only one evaluated.
2 *Ease of comprehension*. The question here for listeners is also general and elicits an overall impression of ease of comprehension. This parameter itself could be further parameterised more objectively by a detailed analysis of what specifically causes problems of comprehension (as in the next feature, for example).
3 *Comprehension problems for individual words*. Here the listener can identify various difficult words, or in a more tightly controlled evaluation experiment the researchers can highlight words known to be difficult and try to analyse the reasons for the difficulties. For example, is there a semantic ambiguity with the word or is it phonologically similar to some other word and insufficiently disambiguated by the semantic context or the syntax?
4 *Intelligibility*. Once again, an overall ranking of general intelligibility.
5 *Pronunciation/occurrence of deviating speech sounds*. Does the listener feel that any particular sounds have been badly rendered and might be contributing to reduced naturalness or acceptability? Notice that errors in sounds in particular combinations will be less noticeable than in other combinations due to the predictability of linear sound combinations in syllables. How frequently do these rogue sounds occur?
6 *Speaking rate*. The question here is whether the speaking rate is appropriate. One difficulty is the extent to which semantic and pragmatic factors enter into the appropriateness of speaking rate. Most synthesisers have a default speaking rate, and maybe should be evaluated on this. Introducing variation of speaking rate may well introduce errors. This is one of the areas–along with other pragmatically sourced variations in

prosody–which will benefit from additional markup of the input text or superior prosody assignment algorithms within the synthesis system.

7 *Voice pleasantness*. A general and very impressionistic parameter, and one which might vary with semantic and prosodic content.

8 *Naturalness*. A general parameter which it may be possible to refine a little. So, we may be able to ask questions like: Is the acoustics of the utterance internally coherent? For example:

- *Does the speech appear to be from a single speaker?*
- *Does this coherence extend throughout the fundamental frequency range with an appropriate amplitude dynamics?*

9 *Liveliness*. In general, liveliness is judged to be a desirable quality contributing to naturalness. But it could be argued that for a general system a whole range of expression along a dullness–liveliness vector should be possible, derived either internally or in response to markup. So the question here is really not

- *Is the speech lively?* but rather
- *Is the degree of liveliness applied to an appropriate degree?*

10 *Friendliness*. This is a quality appropriate for limited domain systems–say, interactive enquiry systems. But in a general system it would be subject, as with naturalness, liveliness and politeness (below), to semantic and pragmatic content. Appropriateness is again a consideration after determining that the default degree of friendliness is convincing.

11 *Politeness*. Again, a subjective evaluation of the default condition–degree of politeness in general–is called for. But also appropriateness for content, and a suitable interpretation of markup, if present, are required judgements.

Each of these parameters is subjective and again defined only vaguely for that reason; and not enough provision is made for adaptation on the part of the listener. But the strength of such an approach is that, notwithstanding the subjective nature of each parameter, the evaluation of naturalness *as a whole* is made more robust. This stems in part from modelling in terms of identifiable features to which a probability might be attached, and in part from the possibility of indicating a relationship between the features. Whilst far from robust in a fully objective way, the characterisation of naturalness here does gain over a non-parametric characterisation, and the approach may eventually lead to productive correlation between measured properties of the soundwave and naturalness. The effects of rendering markup would be an appropriate application for this evaluation technique.

Systems are now good enough for casual listeners to comment not so much on naturalness but on the *appropriateness* of style–as with the friendliness and politeness parameters used by Sluijter *et al*. This does not mean that style, and allied effects, are secondary to naturalness in terms of generating speech synthesis, but it does mean that for some people appropriateness and accuracy of style override some other aspects of naturalness. These considerations would not override intelligibility, which still stands as a prerequisite.

Variability

One of the paradoxes of speech technology is the way in which variability in the speech waveform causes so many problems in the design of automatic speech recognition systems

and at the same time *lack* of it causes a feeling of unnaturalness in synthesised speech. Synthesis seeks to introduce the variability which recognition tries to discard.

Linguistics models variability in terms of a hierarchical arrangement of identifiably different types. We discuss this more fully in Chapter 14, but for the moment we can recognise:

- deliberately introduced and systematic–*phonology*
- unavoidable, but systematic (coarticulation)–*phonetics*
- systematically controlled coarticulation–*cognitive phonetics*
- random–*phonetics*

1 The variability introduced at the phonological level in speech production involves the introduction by the speaker of variants on the underlying segments or prosodic contours. So, for example, English chooses to have two non-distinctive variants of /l/ which can be heard in words like *leaf* and *feel*–classical phonetics called these clear [l] and dark [l] respectively. In the prosody of English we could cite the variant turning-up of the intonation contour before the end of statements as opposed to the usual turn-down. Neither of these variants alters the basic meaning of the utterance, though they can alter pragmatic interpretation. These are termed *extrinsic* variants, and in the segment domain are called *extrinsic allophones*. Failure to reproduce phonological variability correctly in synthetic speech results in a 'foreign accent' effect because different languages derive extrinsic allophones differently; the meaning of the utterance however is not changed, and it usually remains intelligible.
2 Segmental variants introduced unavoidably at the phonetic level are termed *intrinsic allophones* in most contemporary models of phonetics and result from coarticulation. Coarticulation is modelled as the distortion of the intended articulatory configuration associated with a segment–its target–by mechanical or aerodynamic inertial factors which are intrinsic to the speech mechanism and have nothing to do with the linguistics of the language. These inertial effects are systematic and time-governed, and are predictable. Examples from English might be the fronted [k] in a word like *key*, or the dentalised [t] in *eighth*; or vocal cord vibration might get interrupted during intervocalic underlying [+voice] stops or fricatives. Failure to replicate coarticulation correctly in speech synthesis reduces overall intelligibility and contributes very much to lack of naturalness. Interestingly, listeners are not aware of coarticulatory effects in the sense that they cannot report them: they are however extremely sensitive to their omission and to any errors.
3 Observations of coarticulation reveal that it sometimes looks as though coarticulatory effects do vary in a way related to the linguistics of the language, however. The most appropriate model here for our purposes borrows the notion of *cognitive intervention* from bio-psychology to introduce the idea that within certain limits the mechanical constraints can be interfered with–though rarely, if ever, negated completely. Moreover it looks as though some effects intrinsic to the mechanism can actually be enhanced at will for linguistic purposes. Systematic cognitive intervention in the behaviour of the physical mechanism which produces the soundwave is covered by the theory of cognitive phonetics (see Chapter 27). Examples here might be the way coarticulation is reduced in any language when there is a high risk of ambiguity–the speaker slows down to reduce the time-governed constraint–or the enhanced period of vocal cord vibration

failure following some stops in a number of Indian languages. This cognitive intervention to control mechanical constraints enables the enlargement of either the language's extrinsic allophone inventory or even sometimes its underlying segment (phoneme) inventory. If the effects of cognitive intervention in phonetic rendering are not reproduced in synthetic speech there can be perceptual problems occasionally with meaning, and frequently with the coherence of accents within a language. There is also a fair reduction in naturalness.

4 Some random variability is also present in speech articulation. This is due to tolerances in the mechanical and aerodynamic systems: they are insufficiently tight to produce error- or variant-free rendering of the underlying segments (the extrinsic allophones) appearing in the utterance plan. While listeners are not at all sensitive to the detail of random variability in speech, they do become uneasy if this type of variability is not present; so failure to introduce it results in a reduction of naturalness.

Most speech synthesis systems produce results which take into account these types of variability. They do, however, adopt widely differing theoretical stances in how they introduce them. In stand-alone systems this may not matter unless it introduces errors which need not otherwise be there. However, if we attempt to introduce some cross-platform elements to our general synthesis strategy the disparate theoretical foundations may become a problem. In Chapter 20 and Chapter 32 we discuss the introduction of prosodic markup of text input to different synthesis systems. There is potential here for introducing concepts in the markup which may not have been adopted by all the systems it is meant to apply to. A serious cost would be involved if there had to be alternative front ends to copy for different theoretical assumptions in the markup.

Variability is still a major problem in speech synthesis. Linguists are not entirely in agreement as to how to model it, and it may well be that the recognition of the four different types mentioned above rests on too simplistic an approach. Some researchers have claimed that random variability and cognitively controlled intrinsic effects are sometimes avoided or minimised in order to improve intelligibility; this claim is probably false. Cognitive intervention definitely contributes to intelligibility and random variation definitely contributes to naturalness; and intelligibility and naturalness are not entirely decoupled parameters in the perception of speech. It is more likely that some areas of variability are avoided in some synthesis because of a lack of data. In successful state-of-the-art systems, variability is explicitly modelled and introduced in the right places in the planning and rendering algorithms.

The Introduction of Style

Although the quality of text-to-speech systems is improving quite considerably, probably due to the widespread adoption of concatenative or unit selection systems, most of these systems can speak only with one particular style and usually only one particular voice. The usual style adopted, because it is considered to be the most general-purpose, is a relatively neutral version of reading-style speech. What most researchers would like to see is the easy extension of systems to include a range of voices, and also to enable various global styles and local expressive content. All these things are possible–but not yet adopted in systems outside the laboratory.

Prosody control is essential for achieving different styles within the same system. Speech rate control is a good place to start for most researchers because it gives the appearance of being easy. The next parameter to look at might be fundamental frequency or intonation. However, the introduction of speech rate control turns out to be far from simple. The difficulty is expressed clearly in the discovery that a doubling of overall rate is not a halving of the time spent on each segment in an utterance–the distribution of the rate increase is not linear throughout the utterance. Focussing on the syllable as our unit we can see that a change in rate is more likely to affect the vowel nucleus than the surrounding consonants, but it is still hard to be consistent in predicting just how the relative distribution of rate change takes effect. We also observe (Tatham and Morton 2002) that global rate change is not reflected linearly in the next unit up either–the rhythmic unit. A rhythmic unit must have one stressed syllable which begins the unit. Unstressed syllables between stressed ones are fitted into the rhythmic unit, thus:

<utterance>| Pro.so.dy.is | prov.ing | hard.to | mo.del | ac.cur.ate.ly | </utterance>

Here rhythmic unit boundaries are marked with '|' and stressed syllables are underlined. A '.' separates syllables.

The usual model acknowledges the perceptually oriented idea of isochrony between rhythmic units, though despite proving a useful concept in phonological prosody (that is, in the abstract) it is hard to find direct correlates in phonetic prosody–that is, in the actual soundwave. The isochrony approach would hypothesise that the perceived equal timing between stressed syllables–the time between the vertical markers in the above representation–is reflected in the physical signal. The hypothesis has been consistently refuted by researchers.

Evaluating the segmental intelligibility of synthesisers neglects one feature of speech which is universally present–expressive content. In the early days of synthesis the inclusion of anything approaching expression was an unaffordable luxury–it was difficult enough to make the systems segmentally intelligible. Segmental intelligibility, however, is no longer an issue. This means that attention can be directed to evaluating expressive content. In the course of this book we shall return many times to the discussion of expression in synthesis, beginning with examining just what expression *is* in speech. But even if our understanding of expression were complete it would still be difficult to test the intelligibility of synthesised expression. We do not mean here answering questions like 'Does the synthesiser sound happy or angry?' but something of a much more subtle nature. Psychologists have researched the perception of human sourced expression and emotion, but testing and evaluating the success of synthesising expressiveness is something which will have to be left for the future for the moment.

Expressive Content

Most researchers in the area of speech synthesis would agree that the field has its fair share of problems. What we decided when planning this book was that for us there are three major problems which currently stand out as meriting research investment if real headway is to be made in the field *as a whole*. All researchers will have their own areas of interest, but these are our personal choice for attention at the moment. Our feeling is that these three areas

contribute significantly to whether or not synthetic speech is judged to be *natural*–and for us it is an overall improvement in naturalness which will have the greatest returns in the near future. Naturalness therefore forms our *first problem area*.

Naturalness for us hinges on expressive content–expression or the lack of it is what we feel does most to distinguish current speech synthesis systems from natural speech. We shall discuss later

- whether this means that the acoustic signal must more accurately reflect the speaker's expression, or
- whether the acoustic signal must provide the listener with cues for an accurate perceiver assignment of the speaker's *intended* expression.

We shall try to show that these are two quite different things, and that neither is to be neglected. But we shall also show that most current systems are not well set up for handling expressive content. In particular they are not usually able to handle expression on a dynamic basis. We explain why there is a need for dynamic modelling of expression in speech synthesis systems. Dynamic modelling is our *second problem area*.

But the approach would be lacking if we did not at the same time show a way of integrating the disparate parts of a speech synthesis system which have to come together to achieve these goals. And this is our *third problem area*–the transparent integration of levels within synthesis.

The book discusses current work on high-level synthesis, and presents proposals for a unified approach to addressing formal descriptions of high-level manipulation of the low-level synthesis systems, using an XML-based formalism to characterise examples. We feel XML is ideally suited to handling the necessary data structures for synthesis. There were a number of reasons for adopting XML; but mostly we feel that it is an appropriate markup system for

- characterising data structures
- application on multiple platforms.

One of the important things to realise is that modelling speech production in human beings or simulating human speech production using speech synthesis is fundamentally a problem of characterising the data structures involved. There are procedures to be applied to these data structures, of course; but there is much to be gained from making the data structures themselves the focus of the model, making procedures *adjunct* to the focus. This is an approach often adopted for linguistics, and one which we ourselves have used in the SPRUCE model and elsewhere (Tatham and Morton 2003, 2004) with some success.

The multiple-platform issue is not just that XML is interpretable across multiple operating systems or 'in line' within different programming languages, but more importantly that it can be used to manage the high-level aspects of speech synthesis in text-to-speech and other speech synthesis systems which hinge on high-level aspects of synthesis. Thus a high-level approach can be built which can precede multiple low-level systems. It is in this particular sense that we are concerned with the application across multiple platforms. As an example we can cite the SPRUCE system which is essentially a high-level synthesis system whose output is capable of being rendered on multiple low-level systems–such as formant-based synthesisers or those based on concatenated waveforms.

Final Introductory Remarks

The feeling we shall be trying to create in this book is one of optimism. Gradually inroads are being made in the fields of speech synthesis (and the allied field of automatic speech recognition) which are leading to a greater understanding of just how complex human speech and its perception are. The focus of researchers' efforts is shifting toward what we might call the *humanness* of human speech–toward gaining insights into not so much how the message is encoded using a small set of sounds (an abstract idea), but how the message is cloaked in a multi-layered wrapper involving continuous adjustment of detail to satisfy a finely balanced interplay between speaker and listener. This interplay is most apparent in dialogue, where on occasion its importance exceeds even that of any messages exchanged. Computer-based dialogue will undoubtedly come to play a significant role in our lives in the not too distant future. If conversations with computers are to be at all successful we will need to look much more closely at those areas of speaker/listener interaction which are beginning to emerge as the new focal points for speech technology research.

We have divided the book into a number of parts concentrating on different aspects of speech synthesis. There are a number of recurrent themes: sometimes these occur briefly, sometimes in detail–but each time from a different perspective. The idea here is to try to present an integrated view, but from different areas of importance. We make a number of proposals about approach, modelling, the characterisation of data structures, and one or two other areas: but these have only the status of suggestions for future work. Part of our task has been to try to draw out an understanding of why it is taking so long to achieve genuinely usable synthetic speech, and to offer our views on how work might proceed.

So we start in Part I with establishing a firm distinction between high- and low-level synthesis, moving toward characterising naturalness as a new focus for synthesis in Part II. We make suggestions for handling high-level control in Part III, highlighting areas for improvement in Part IV. Much research has been devoted recently to markup of text as a way of improving detail in synthetic speech: we concentrate in Part V on highlighting the main important advances, indicating in Part VI some detail of how data structures might be handled from both static and dynamic perspectives. A key to naturalness lies in good handling of prosody, and in Part VIII we move on to some of the details involving in coding and rendering, particularly of intonation. We present simple ways of handling data characterisation in XML markup, and its subsequent processing with examples in procedural pseudo-code designed to suggest how the various strands of information which wrap the final signal might come together. Part IX pulls the discussion together, and the book ends with a concluding overview where we highlight aspects of speech synthesis for development and improvement.