# A Phonetically Transparent Technique for the Automatic Transcription of Speech

Michael J. Morony

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
to the
University of Edinburgh
1998

# Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

April 1998

# Abstract

This thesis describes a technique for automatic transcription of speech that is both probabilistic and phonetically transparent. The technique employs classical methods of statistical pattern-recognition, but works within a phonetic framework that is simple to conceptualise, with no "black box" components or hidden states or layers.

The technique being statistical, it involves a prior training-phase during which the features of the classes to be recognised are learnt from examples, and statistical models estimated for each class. These models subsequently become the basis of the recognition procedure. In the present implementation, the classes are classes of "sub-phonic" entities (where "sub-phonic" denotes parts of phones and phones are conceived as artificially isolated sections of the acoustic record of speech, each phone serving an identifiable linguistic function). The sub-phonic entities (or 'subphones') are defined in such a way as to take account of the underlying articulatory reality, and the manner of their definition imposes a significant amount of constraint on the search for the most probable utterance-transcription, a search executed using dynamic programming.

The subphones are identified explicitly and modelled statistically in their own right. Explicit identification and modelling result in greater simplicity than is typical of other statistical modelling techniques. No use is made of durational features in their identification either in training or (in the basic implementation of the technique) in recognition. It is argued that duration should be modelled primarily not as a feature of phonetic classes, but rather of higher level structures, though it is also suggested that subphonic analysis of phones may provide a basis for relativistic within-phone duration-modelling.

The technique described here is offered as a possible starting-point for further development. Evaluation of the technique suggests that it is realistic to claim that substantial improvement is possible from the current best performance for a single speaker of a little under 70 per cent correct for phone recognition (this without syllabic, lexical, syntactic or other higher-level constraints, and without the use of mixture-modelling). Because the phonetic framework is transparent, it is considered probable that rapid progress

would be possible from this starting-point using the tried and tested methods of empirical science.

# Dedication

This work is dedicated to the memory of my father, Frank Morony (1922 – 1952).

# Acknowledgements

I would like to express my thanks to my supervisor, Steve Isard, for all his help over the five years this thesis took to prepare. No one could have been more generous with their time, and I benefited enormously from his wide knowledge, capacity for clear thinking, and good sense. I also wish to express my very warm thanks to Fergus McInnes, who kindly agreed to act as my second supervisor, and who was always extraordinarily generous in the care and detail with which he answered my questions and read through early drafts. He made numerous suggestions and criticisms which resulted in significant improvements both to the thesis and to the system it describes. I feel that I was very privileged in having access to his advice.

I owe a debt to all the computing staff in the Department of Linguistics. I would particularly like to thank Norman Dryden for all his help and advice over a wide range of computing topics, and also for the encouragement he gave me when my morale was low. A special word of thanks is due also to Cedric McMartin for helping me out of programming holes on a number of occasions.

Others who have helped me in significant ways include Art Blokland, Yuko Kondo, Sally Bates, Paul Taylor of CSTR, Diego Molla, Geoff Lindsay, and Keith Edwards of CCIR, and I would like to express my thanks to them all. Thanks finally to Professor Jim Hurford, who first encouraged me to consider embarking on a Ph.D., and whose kindness and encouragement were very much appreciated.

# Abbreviations used in the Text

| Abbreviation | Meaning |
|---|---|
| APT | automatic phonetic transcription |
| ASR | automatic speech recognition |
| AST | automatic subphonic transcription |
| CSBCM | comprehensive separation between classes and models |
| DCT | Discrete Cosine Transform |
| DP | Dynamic Programming |
| GSW | initials of the speaker who provided the speech data |
| HMM | Hidden Markov Model/Modelling |
| HSMM | Hidden Semi Markov Model/Modelling |
| OQ | Open Quotient |
| pdf | probability density function |
| PSSH | the Phonetic Self-Sufficiency Hypothesis |
| POA | place of articulation |
| RD | relative duration |
| SBCM | separation between classes and models |
| TR Classes | classes used to model the border between two phones |
| FURIDA | the automatic transcription system described in this work |
| VTC | vocal tract configuration |

# Contents


**1 Preliminaries** 1

1.1 Introduction . . . . . 1
1.2 Recognition, Perception and Classification . . . . . 2
1.3 Recognition of Language . . . . . 5
1.4 "The Phonetic Self-Sufficiency Hypothesis" . . . . . 7
1.5 The Speech Material Used in Training and Testing the Recognition System 9
1.6 Brief Preview of the Content of Subsequent Chapters . . . . . 9

**2 The Phonetic Framework** 11

2.1 Introduction . . . . . 11
2.2 Basics of the Statistical Approach to ASR . . . . . 12
2.3 ASR and Linear Segmental Models of Speech . . . . . 13
2.4 Brief System Overview . . . . . 17
2.5 Segmental Units, Sequence-Constraints and Timing-Variation . . . . . 20
2.5.1 Fricatives . . . . . 22
2.5.2 Stops . . . . . 37
2.5.3 Affricates . . . . . 40
2.5.4 Nasals . . . . . 41
2.5.5 Laterals . . . . . 57
2.5.6 [r] and [w] . . . . . 68
2.5.7 [h] . . . . . 75
2.5.8 [y] . . . . . 78
2.5.9 Monophthongal Vowels . . . . . 79
2.5.10 Diphthongs . . . . . 82
2.6 Elision, Assimilation and Fusion . . . . . 86
2.7 Glottalisation – the Segmental Framework tested to the limit . . . . . 89
2.7.1 Clear Cases of Stop Substitution . . . . . 90
2.7.2 Stop Marking of Doubtful Species . . . . . 95

2.7.3 Juncture Emphasis . . . . . 99
Full Juncture Glottalisation . . . . . 100
Weak Juncture Glottalisation . . . . . 101

**3 Representing the Short-term Spectrum 105**
3.1 Introduction . . . . . 105
3.2 The Short-term View . . . . . 106
3.3 Features Desirable in a Representation Vector . . . . . 108
3.4 Mel Cepstral Representations . . . . . 108
3.4.1 Extraction of the Most Salient Features of the Spectrum . . . . . 109
3.4.2 Separation of Contributions from Source and Filter . . . . . 113
3.4.3 Frequency-Warping and Critical Bands . . . . . 115
3.5 Window-Size and Stationarity – Border-Straddling ('TR') Classes . . . . 130
3.6 Incorporating Trend Information . . . . . 133
3.7 Summary of the Representation Used . . . . . 138

**4 Getting from Cepstra to Phones 139**
4.1 Introduction . . . . . 139
4.2 Training and Recognition Procedures in Outline . . . . . 139
4.3 Statistical Pattern Classification . . . . . 142
4.4 Dynamic Programming . . . . . 146
4.5 Further Details of the Training Procedure . . . . . 151
4.5.1 The Initialisation . . . . . 151
4.5.2 Iterative Closed-Test Reclassification . . . . . 155
4.6 Further details of the Recognition Procedure . . . . . 159
4.6.1 Considerations of Efficiency . . . . . 159
4.6.2 TR classes in the Transcription Procedure . . . . . 162
4.6.3 Stop Closures in the Transcription Procedure . . . . . 170
4.6.4 Converting from the Subphonic Transcription to A Phonetic Transcription . . . . . 175
4.7 A Comparison with Hidden Markov Modelling (HMM) . . . . . 176
4.7.1 Introduction . . . . . 176
4.7.2 Training . . . . . 179
4.7.3 Duration Modelling in HMM and in Hidden Semi-Markov Modelling 182
4.8 Incorporating Duration-Modelling in FURIDA . . . . . 186
4.8.1 Where Does the Lack of Duration Modelling Tell? . . . . . 186
4.8.2 Some Questions Regarding The Mechanics of Duration Scoring . 190

4.9 Conclusion and Summary . . . . . 192

**5 Making the Most of Limited Training Data 194**

5.1 Introduction . . . . . 194
5.2 Minimum Requirements for Statistical Modelling . . . . . 196
5.3 Generalisation of Classes in FURIDA . . . . . 197
5.3.1 Basic Principles of Generalisation . . . . . 197
5.3.2 Mechanics of Conditional Generalisation . . . . . 204
5.3.3 A Closer Look at Selected Trees . . . . . 209
5.3.4 Problems of Bias . . . . . 213
5.3.5 Defining Predecessor-Sets to Accommodate Generalised Classes . 214
5.4 Standard Approaches to Clustering . . . . . 216
5.4.1 Agglomerative Clustering . . . . . 217
5.4.2 Clustering Using Automatically Induced Decision-Trees . . . . . 219
5.5 Further Options for Dealing with Data-Shortages . . . . . 222
5.5.1 Review of Elementary Statistical Properties of Multivariate Sample-Data . . . . . 222
5.5.2 Statistical Profile of the Class Sample Data . . . . . 225
5.5.3 Model-Fitting, Training Quotas, and Simplifying Assumptions . 229
5.5.4 Pooling and Thresholding of Variance Estimates for Small Samples 231

**6 Evaluation and Experimental Results 233**

6.1 Introduction . . . . . 233
6.2 The handling of "TR" classes, Part I . . . . . 242
6.3 Completeness and Effectiveness of Sequence Constraints . . . . . 244
6.4 Cost to the system of the lack of any explicit duration modelling . . . . 250
6.5 Relative Duration Penalties — The Problem of "Gliditis" . . . . . 253
6.5.1 The problem . . . . . 253
6.5.2 Possible causes . . . . . 254
6.5.3 Possible solutions . . . . . 254
6.5.4 The Experiments . . . . . 255
6.5.5 Frame-level Discriminability for the Phones Involved . . . . . 261
6.6 Effects of the assumption of Normal distributions . . . . . 262
6.7 The handling of "TR" classes, Part II . . . . . 276
6.8 Representativeness of Training Data . . . . . 281
6.9 Proportion of Error Lying in Pre-Pausal Syllables . . . . . 290
6.10 Cepstral Representations . . . . . 292

6.11 Concluding Remarks . . . . . 298

**7 Conclusion** **299**

**A Phone Labels** **304**

**B HResults Aligned Transcriptions** **309**

**C TR Classes** **333**

# List of Figures


2.1 Transcription of Subphones at Phone Border Between [x] and [y] . . . . 19
2.2 [s ir] from utterance of "considered" (sc003) . . . . . 24
2.3 [s uuf] from utterance of "soon" (sc059) . . . . . 25
2.4 [s e] from utterance of "several" (sc018) . . . . . 26
2.5 [s a] from utterance of "sandwiches" (sc052) . . . . . 27
2.6 [s ax] from utterance of "habits of" (sc008) . . . . . 28
2.7 [s ax] from utterance of "since" (sc017) . . . . . 29
2.8 [s ai] from utterance of "decidedly" (sc105) . . . . . 30
2.9 [s] (on left) and [s T] from utterance sc104 (see text) illustrating the difficulty of [s] vs [s T] discrimination . . . . . 32
2.10 [i n ir s] from utterance of "minister" (sc008) . . . . . 43
2.11 [ax m a] from utterance of "a magazine" (sc020) . . . . . 44
2.12 [aa n] from utterance of "Blanche" (sc107) . . . . . 45
2.13 [e n i] from utterance of "Jennings" (sc107) . . . . . 46
2.14 [aD1 n e2ii] from utterance of "our neighbours" (sc118) (where [aD1] represents a realisation of the diphthong [au] as in "now" . . . . . 47
2.15 [n f] from utterance of "own photographs" (sc114) . . . . . 49
2.16 [n f] from utterance of "can finish" (sc127) . . . . . 50
2.17 [n s] from utterance of "has been served" (sc146) . . . . . 51
2.18 [n kc] from utterance of "an' cranny" (sc115) . . . . . 52
2.19 [aD1 n kc] from utterance of "ground coffee" (sc004) . . . . . 53
2.20 [t e GLkc NPn o] from utterance of "technology" (sc005), (where [GLkc] is a closure for a [k] that is preceded by glottalisation) . . . . . 55
2.21 [k ax l a] from utterance of "collapsed" (sc150) . . . . . 60
2.22 [ei cl oo] from utterance of "they launched" (sc011) . . . . . 61
2.23 [t uuf l ax th] from utterance of "too lethargic" (sc168) . . . . . 62
2.24 [s lo l e] from utterance of "slept" (sc168) . . . . . 64
2.25 [s lo ax s] from utterance of "tasteless" (sc185) . . . . . 65

2.26 [s lo l ii] from utterance of "sleep" (sc187) . . . . . 66
2.27 [trc trb r1 r2 ei] from utterance of "Mediterranean" (sc093) . . . . . 71
2.28 [trc trb r2 i] from utterance of "loitering" (sc008) . . . . . 72
2.29 [trc trb aa n] from utterance of "entrancing" (sc160) . . . . . 73
2.30 [a GLpc pb s] from utterance of "collapsed" (sc150) . . . . . 91
2.31 [e GLpc pb th] from utterance of "depth" (sc111) . . . . . 92
2.32 [a GLpc f] from utterance of "von Trapp family" (sc122) . . . . . 93
2.33 Forms/Functions of Glottalisation . . . . . 93
2.34 [kc krb r2 e dnc ir GLkc kb aa d] ("credit card") . . . . . 97
2.35 [ax v GLuuf z] from utterance of "have oozed" (sc122) . . . . . 102
2.36 [ou PVdh GLaaD1 aa2i iD3] from utterance of "loathe eiderdowns" (sc122) 103

3.1 First Four Basis Functions of DCT, showing transform coefficients on the y-axis for values of x representing ordinal points in the series of FFT spectral coefficients . . . . . 110
3.2 Fifth to Eighth Basis Functions of DCT, axes as in previous figure . . . 111
3.3 [a] vs [e] under scheme A . . . . . 122
3.4 [a] vs [e] under scheme B . . . . . 123
3.5 [a] vs [e] under scheme C . . . . . 124
3.6 [a] vs [e] under scheme D . . . . . 125
3.7 [ii] vs [uuf] under scheme A . . . . . 126
3.8 [ii] vs [uuf] under scheme B . . . . . 127
3.9 [ii] vs [uuf] under scheme C . . . . . 128
3.10 [ii] vs [uuf] under scheme D . . . . . 129
3.11 TR31: [tc]/[tb] borders . . . . . 133
3.12 TR32: [dc]/[db] borders . . . . . 134
3.13 TR33: [bc]/[bb] borders . . . . . 134
3.14 TR34: [pc]/[pb] and [pc]/[Pb] borders . . . . . 135
3.15 TR35: [gc]/[gb] borders . . . . . 135
3.16 TR36: [kc]/[kb] and [kc][Kb]borders . . . . . 136

4.1 Training and Recognition Procedures in Outline . . . . . 140
4.2 Linkage Across TR 'Segments' . . . . . 164
4.3 Old Scheme for Remote Sequence Constraints . . . . . 165
4.4 New Scheme for Remote Sequence Constraints . . . . . 166
4.5 Allowing Optimal Placement of TR 'Segments' (I) . . . . . 168
4.6 Allowing Optimal Placement of TR 'Segments' (II) . . . . . 169
4.7 Application of Stop-closure Duration-Penalties . . . . . 174

5.1 Levels in a Tree . . . . . 206
5.2 Downward Propagation of Marks . . . . . 208
5.3 Generalisation Tree Schema for Vocalic Right Contexts of Offset-Subphones of Fricatives, Laterals and Nasals . . . . . 210
5.4 Generalisation Tree Schema for Vocoid Left Contexts of Onset-Subphones of Fricatives, Affricate-Releases, Laterals and /y/ . . . . . 211
5.5 Generalisation Tree Schema for Consonantal Left Contexts of Onset-Subphones of Dark Laterals . . . . . 212
5.6 Plot of 1st static against 1st delta coefficient for PAL_MHBB . . . . . 224
5.7 Dispersion of Correlations Across All Classes . . . . . 228

6.1 two normal populations treated as one . . . . . 263
6.2 Static Cepstral Coefficients for [Caa] . . . . . 264
6.3 Static Cepstral Coefficients for [Cii] . . . . . 265
6.4 Static Cepstral Coefficients for [Coo] . . . . . 265
6.5 Static Cepstral Coefficients for [Cuuf] . . . . . 267
6.6 Static Cepstral Coefficients for [Cuum] . . . . . 267
6.7 Static Cepstral Coefficients for [Cuu] . . . . . 268
6.8 Static Cepstral Coefficients for [Crii] . . . . . 269
6.9 Static Cepstral Coefficients for [Ci] . . . . . 269
6.10 Static Cepstral Coefficients for [HF2_VELA] . . . . . 270
6.11 Static Cepstral Coefficients for [R_IB] . . . . . 270
6.12 Static Cepstral Coefficients for [R_RIIB] . . . . . 271
6.13 Static Cepstral Coefficients for [AA_NVCORA] . . . . . 272
6.14 Static Cepstral Coefficients for [AA_VCORA] . . . . . 272
6.15 Static Cepstral Coefficients for [HB2_VCORA] . . . . . 273
6.16 Static Cepstral Coefficients for [kb2_AA2A] . . . . . 274
6.17 Static Cepstral Coefficients for [kb2_MHBA] . . . . . 274
6.18 Static Cepstral Coefficients for [TR10] . . . . . 275
6.19 Static Cepstral Coefficients for [TR60] . . . . . 276
6.20 Training Data (static coefficients) for [IIY_zB] . . . . . 283
6.21 Test token (static coefficients) [IIY_zB] . . . . . 284
6.22 Training Data (dynamic coefficients) for [IIY_zB] . . . . . 285
6.23 Test token (dynamic coefficients) [IIY_zB] . . . . . 285
6.24 Combined plot of Training and Test token data (dynamic coefficients) for [IIY_zB] . . . . . 286
6.25 Training Data (static coefficients) for [E2_NVCORA] . . . . . 287

6.26 Test token (static coefficients) [E2_NVCORA] . . . . . 287
6.27 Training Data (dynamic coefficients) for [E2_NVCORA] . . . . . 288
6.28 Test token (dynamic coefficients) [E2_NVCORA] . . . . . 288
6.29 Combined plot of Training and Test-token data (dynamic coefficients) for [E2_NVCORA] . . . . . 289

# List of Tables


2.1 Fricative Labels . . . . . 22
2.2 Fricative Allophones . . . . . 36
2.3 Context-Dependency of Fricative Phones and Subphones . . . . . 36
2.4 Stop Labels . . . . . 37
2.5 Conversion of Specific Stop Closure Labels to Generic Labels . . . . . 38
2.6 Context-Dependency of Stop Phones and Subphones . . . . . 41
2.7 Nasal Labels . . . . . 41
2.8 Context-Dependency of Nasal Phones and Subphones . . . . . 57
2.9 Context-Dependency of Lateral Phones and Subphones . . . . . 69
2.10 Labels for Monophthongal Vowels . . . . . 79
2.11 Two-Phase Diphthongs . . . . . 82
2.12 Manual Labels for Maximal Realisations of Rising Diphthongs . . . . . 83
2.13 Automatic conversion of post-consonantal glides . . . . . 84
2.14 Conversion of post-D1 glide labels . . . . . 85
2.15 Automatic Conversion of Glottalisation Labels . . . . . 94
2.16 Automatic Conversion of Glottalisation Labels . . . . . 98

3.1 Critical Band Specifications . . . . . 117

4.1 Matrix of Class/Frame Scores for Five-Frame Vowel . . . . . 150
4.2 Accumulated Best Sub-Path Scores, and Best Predecessors in Brackets . 151
4.3 Stop Duration Penalties . . . . . 172

5.1 Immediate Merges of Vowel Subphones . . . . . 198
5.2 Immediate Merges of Consonantal Contexts of Vowels . . . . . 201
5.3 Immediate Merges of Consonantal Contexts of Vowels (continued) . . . 202
5.4 Dispersion of Variances Across All Classes . . . . . 227

6.1 Monolithic TR Segments . . . . . 243
6.2 Nominal Resolution of TR Segments . . . . . 243

6.3 Confidence and Significance Measures . . . . . 243
6.4 Nominal Resolution of TR Segments . . . . . 246
6.5 With Additional Sequence-Constraints . . . . . 246
6.6 Confidence and Significance Measures . . . . . 246
6.7 No Duration Penalties . . . . . 252
6.8 With Duration Penalties Added . . . . . 252
6.9 Confidence and Significance Measures . . . . . 252
6.10 Ratios and Penalties for Diphthongs with glide to [ax] . . . . . 257
6.11 Ratios and Penalties for Diphthongs with glide to [ii], [i] or [e] . . . . . 258
6.12 Absolute Duration Penalties only . . . . . 259
6.13 With Relative Duration Penalties added . . . . . 259
6.14 Confidence and Significance Measures . . . . . 259
6.15 Absolute Duration Penalties only, 190 Training Sentences . . . . . 260
6.16 Relative Duration Penalties added, 190 Training Sentences . . . . . 260
6.17 Parameters for data-sets A, B and C = A+B . . . . . 262
6.18 With Relative Duration Penalties, Nominal Resolution of TR's . . . . . 277
6.19 With Relative Duration Penalties, Real Resolution of TR's (method 1(a)) 278
6.20 Confidence and Significance Measures . . . . . 278
6.21 Real Resolution of TR's, method 1(b) . . . . . 279
6.22 Confidence and Significance Measures . . . . . 279
6.23 Real Resolution of TR's, method 2 . . . . . 280
6.24 Confidence and Significance Measures . . . . . 280
6.25 Closed Test Results . . . . . 281
6.26 Results for Recognition of Final Four Phones . . . . . 290
6.27 Results for Recognition of Final Four Phones, without use of deltas . . . 291
6.28 Complete Transcription as Reference, with no use of deltas . . . . . 292
6.29 R26 Banding Scheme . . . . . 293
6.30 R26 Results . . . . . 293
6.31 Bark Representation Results . . . . . 293
6.32 Confidence and Significance Measures . . . . . 294
6.33 R24 Banding Scheme . . . . . 294
6.34 R24 Results . . . . . 294
6.35 R31 Banding Scheme . . . . . 295
6.36 R31 Results . . . . . 295
6.37 R23 Banding Scheme . . . . . 296
6.38 R23 Results . . . . . 296
6.39 TRI33 Banding Scheme . . . . . 297

6.40 TRI33 Results . . . . . 297
6.41 Confidence and Significance Measures . . . . . 297

# Chapter 1

# Preliminaries

## 1.1 Introduction

The work on which this thesis is based was directed to an essentially practical end – the development of a phonetically transparent technique for automatic transcription of speech. By "phonetically transparent" is meant a technique whose manner of operation is wholly intelligible in phonetic terms; a fuller understanding of the significance of this will emerge as the thesis unfolds. In order to keep the task within reasonable bounds, the attempt was restricted to the single-speaker case (with the idea, however, that if success could be achieved for a single speaker, extension to the case of several speakers could be undertaken as a further step). A more minor restriction was imposed by the decision to terminate with a phonetic rather than a full orthographic transcription (a restriction described as minor in view of the fact that a conventional orthographic transcription should be derivable in a fairly straightforward way from the phonetic transcription produced).

To a large extent, the success achieved to date in automatic speech recognition has been proportional to the degree to which the tasks attempted have been constrained to involve a limited number of choices at each decision-point. The task of automatic speech *recognition* (ASR) is distinguished from that of automatic speech *understanding*, and it is probably true to say that the artificial constraints imposed on the tasks attempted in ASR serve to compensate for the loss of an 'overseeing intelligence'. Certainly in human language perception, understanding

appears to play a significant and probably quite crucial role in enabling us to recognise what is said. The separation of recognition and understanding reflects only an attempt to tackle a difficult problem by dividing it into less complex problems, each of which is difficult enough by itself. The presence of artificial constraints in a typical ASR task — for example, using a finite lexicon and limited grammar to define the range of things that can be said — may obscure the extent to which the purely phonetic characterisation of speech is being well managed, and it is possible to forgo such constraints in order to get a clearer picture of the phonetic sensitivity or performance of a technique (Lamel & J.L.Gauvain 1993). This is what was attempted in the work described here.

It cannot escape notice, however, that in the human context language works within limitations imposed by a finite (if large) lexicon, a definite (if rich) grammar, and a generally accessible and familiar world of reference, so to speak. It is therefore far from obvious that phonetic transcription *should* be possible without assistance from the other possible sources of help. Most of the rest of this thesis is concerned in one way or another with the practicalities of the attempt, but in this opening chapter I take time to consider whether the idea is even sensible. I begin with a brief examination of the concepts of recognition and perception, first as they apply to human experience in general, and then as they apply to our experience of language in particular.

## 1.2 Recognition, Perception and Classification

In ordinary usage, *recognition* is most often recognition of individuals (in the philosophical sense of things existing as entities in their own right) as the unique individuals they are — human individuals, places, people's handwriting or pets or material belongings, or other inanimate objects. It is easy to catalogue the kinds of things we talk of recognising in this way, but not so easy — as with all terms involving processes that seem to take place within us — to say what recognition consists in, unless we give a rather obvious definition as, say, coming to realise that the individual before us (in flesh, film, or whatever) is one that we know, that we have had previous experience of. The definition is not very enlightening, but it does at least serve to draw attention to the central importance

for all recognition, and probably for perception in general, of a stored memory or mental construct of some kind which has been built from previous experience of the individual in question (Kohonen 1989:p. 1).

It is less common to describe perception in the general sense — our everyday perception of the things in our environment, for example — in terms of recognition, but I think most people would allow that in some sense we have to recognise the things in our environment — as cars, shop-doorways, ten-pound notes, policemen — in order to live our lives at all. In this sense recognition is a matter of seeing what something is, often with the emphasis not on its unique individuality (*has* this ten-pound note passed through my hands before? *What does it matter?*) but on its nature or value or role as a thing of that kind. And once again, *experience* of other things of that kind is crucial in having enabled us to develop a mental construct which allows us to see this thing too as a thing of that kind, even though we may have had no previous experience of it whatsoever *qua* individual, and so to know how to behave with regard to it (if it's a newspaper, we know we can read it; if it's a moving car, we know we should not stand in its way).

In so far as perception of everyday objects involves recognising them as belonging to particular kinds, it can reasonably be said to involve acts of *classification* [1], even if the word is rather too formal to make the locution completely natural. If asked to explain the term *classification* to a non-native learner of English, we would probably use as examples either formal, scientific classification involving specialised knowledge, or conceivably (stretching things a bit) cases where some unfamiliar object — a particular small metal item found in a disused factory, say, or some barnacle-encrusted object found in a rock-pool — required to be identified – as a battered rivet, say, or a child's seaside spade. Classification of the formal or explicit kind does not appear to be very different in the most important respects from recognition of things in the course of everyday perception, the major differences arising only from the degree of ease with which the process is carried out in the different cases, this probably being a function of familiarity more than anything else. While a careful process of taking measurements, noting

[1] in the sense, clearly, of putting an item into a class, rather than of devising a taxonomy or classification-scheme for performing such acts

down distinctive features, and the like, may not be required in these less formal acts of classification, it seems incontrovertible to say that we can only classify or recognise things by means of their features or properties, though the latter terms should surely be interpreted very widely to include networks of relationships with other things.

It is surely wrong to think that we rely *exclusively* on properties of things themselves in recognising them — our memory or mental picture of, for example, the world of travelling to work in the morning, with all the expectations it creates as we travel to work *this* morning, means that only the most cursory attention to the features of a moving object in the road will be sufficient for us to recognise it as a car. This is not, of course, to deny the importance of the car's having the features that enable us to see it as a car, but only to draw attention to the very great power of a known context to facilitate recognition of things that we would expect to find in that context.

For a sane human being, the inventory of things in their mental construct of the world will reflect the inventory of things in their experience of the world, and will predispose them to expect their future experience of the world to be at least in large part concerned with items from that same inventory. In trying to recognise an obscure object, indeed, we may try out a number of ideas (mental constructs, patterns), and often be able to arrive at what seems like a plausible recognition on this basis. Recognition is thus shown to be an active process of "reconstruction", at least in the cases where the relating of object-of-perception and object-of-experience (mental construct) is not straightforward, and one is perhaps entitled to suspect therefore that it is everywhere an active process, only that it is usually so easy as to be for all practical purposes "automatic", and so passes unnoticed.

We are not passive in perception, opening blank senses to have a pre-determined world impinge upon them and give rise to a mental construct of the world (or a mental construct of mental constructs (...) of the world) that could in some sense be described as objective. As humans, we were initiated into the world of human experience in our infancy, and the nature of our humanity, with the relationships this involves us in vis-a-vis the rest of the world, human and non-human, surely determines how the world is "divided up" for us, and what things

become "objects of experience" for us. And needless to say, those things become objects of experience for us which are significant to our lives. At any one moment, however, our resolution of the world before us is determined by the manner of our engagement with it at that time, with some things backgrounded and some things brought into primary focus. We may, at one moment, pay very close attention to the features of an individual (consider the lepidopterist examining a rare butterfly), and at another merely notice the individual as in a sense no more than a feature of a grander whole, as we might do with respect to a butterfly passing across the garden on a Summer's day. The "objects of significance", therefore, may be different in scale, so to speak, from one occasion to another.

## 1.3 Recognition of Language

Whether or not *everything* that has been said above about the world of our experience can be said also of language as a particular case, it certainly seems that our capacity to learn and to use language reflects many of the same features of our general capacity for learning about and making sense of the world. We recognise utterances of word X as belonging to the kind "utterance of X" on the basis of their having properties (sound-patterns) that we have learnt to be associated with utterances of X, and properties (patterns of use, involving both connections with the world and connections with other words) that we have learnt to be part of the identity of the word X. Every word, that is to say, has both a spoken form and a meaning or use, and while its spoken form is vital as a key to its identity, its use in the language creates for it a network of relationships that are surely also significant in facilitating its recognition. (It does seem to be correct to speak of "*the* word X", to give the type-word the status of a unique individual, while treating utterances of the word as forming a class.)

There is little doubt that we internalise the landscape of our language just as strongly as we internalise our learnt picture of the world, and we appear to use our knowledge of both to make sense of what we hear. As we begin to recognise the words of an utterance, patterns of implication and patterns of association are invoked with respect both to the linguistic form of the utterance and to its semantic interpretation, and may exercise a significant influence on our decision

as to what we are hearing. And just as it is with aspects of the world that the more familiar we are with them the less care and attention we have to expend when dealing with them, so too does it seem to be the case in normal circumstances with language, that the more experienced we are in its use the less detail we need in order to be be able to make sense of it, a fact which we betray our implicit knowledge of in the way we adapt our speaking style with a hearer who we know to be familiar with our accent, our speech habits, or the subject-matter of our talk. In the extreme case, the phonetic delineation of the linguistic content of an utterance may be merely sketched, on the assumption that other factors of the situation make it almost unnecessary to formulate the utterance at all.

The parallel between language-perception and perception of the world in general would appear to extend to the points made about "objects of significance" also. The words of our language are at the very least *one* of the categories of "object of significance" for us as we participate in language (many would argue that they are the primary category), while the phrases and other structures formed from words may also be capable of being objects of significance in this sense (consider phrases such as "How are you?", "I love you", "We're ruined!", "What's the meaning of this?"). Whether or in what sense the sounds used to *form* words are themselves objects of significance for the naive language-user is a question whose correct answer is somewhat uncertain. It seems true to say that we do not learn our mother-tongue by learning to produce or recognise sounds in their own right, but rather are taught to produce and recognise words, which involve coordinated articulatory gestures for their production and involve recognition of changing patterns of sound, but which seem to be produced and recognised as wholes (Hawkins 1995). On the other hand, a variety of factors conspire to make children gradually more aware of similarities between the sounds involved in different words, and in societies with alphabetic writing-systems a person may be led to assume from the fact that a single symbol is used for a group of very similar sounds (take the vowel-sounds in the words *bit*, *sit*, *hid* and *quick* for example), that they are in some sense all *the same* sound. This is not strictly true, of course, of the *physical* sounds, but a significant point about assuming something like it is that making such assumptions about the sound-patterns of languages has been found by humankind over a long period of time to lead to useful results

in a number of practical applications (user-friendly writing-systems; guides to pronunciation in dictionaries for languages like English with its very awkward relationship between spelling and sound, or for languages like Chinese with no systematic connection between the written and spoken forms of the language; as a starting-point for pronunciation teaching in foreign and second language teaching, ... ). It is still, I think, worth insisting that the fact that we can focus on phonetic or phonemic classes of sounds does not entail that we actually engage in such activity, or must engage in such activity, when listening to speech. It is perhaps worth adding also that whether we do or not, it is unlikely in the extreme that we make sense of language *simply* by means of such a process (that we work entirely "bottom-up" from the perception of individual sounds).

## 1.4 "The Phonetic Self-Sufficiency Hypothesis"

It is a basic (and I think, uncontroversial) assumption of this thesis that it is possible to *represent* the linguistic content (the "message") of at least a carefully produced utterance using a finite set of phonetic symbols *once the linguistic content is known*. The more questionable assumption is that the linguistic content of an unknown utterance can be recovered from its bare acoustic record, given a list of the phonetic categories to be employed in the recovery and a known mapping from categories to physical sounds. I am calling this the phonetic self-sufficiency hypothesis (PSSH) for the rest of this section. Stating the matter in rather more detail, the assumptions involved are that, for a given speaker, (a) speech can be seen as involving the recurrent use of a finite number of sounds, (b) the mapping of classes of sound to distinctive categories is learnable, and (c) once the mapping has been learnt, the acoustic record of that speaker's utterances should be sufficient for recovery of a phonetic representation, without any need to rely on lexical or grammatical or other higher-level constraints to constrain the transcription process. This may be described as the strong form of the hypothesis. A weaker form states that a *probabilistic* phonetic representation may be obtained as above, where "probabilistic" implies a number of different possible transcriptions each comprising scored phonetic hypotheses, from which a definitive phonemic transcription is achievable using a lexicon and grammar and

possibly other higher-level constraints. (Note that under the weak form of the hypothesis at least, the view is implied that the act of "recognising a phoneme" is only completed when recognition of the word it helps to identify is complete.)

That the weak form (at least) of the hypothesis is at least not obviously false may be argued as follows. Language in its spoken form is only possible if words have *some* degree of stability in their spoken forms (variation must have its limits), which implies a learnable relationship between any particular word and pronunciations of that word in the speech of a given speaker, and implies also the maintenance of a more or less fixed system of relationships of this kind across the entire lexicon (so that, for example, if the speaker produces the diphthong of 'my' as something akin to [oi], as many Dubliners do, he may be expected to use the same sound [oi] in the words 'sky', 'mine', 'high' and 'mind'). (If people continually changed their phonetic keys to particular distinctive categories, it would probably be quite difficult to follow them.) It therefore seems not unreasonable to expect that once we learn the system of relationships, it will be possible to recognise the phonetic content of further utterances by that speaker from their acoustic record (the strong form of PSSH), or at least have evidence which we can use in combination with help from other sources to arrive at the linguistic content (the weak form of the hypothesis).

This optimistic view of the adequacy of the acoustic-phonetic record may be thought to sit ill with the view argued for earlier that a variety of knowledge-sources contribute to facilitating human recognition of language. Two points are relevant in this connection. Firstly, in the real world (with multitudes of speakers) there is a huge amount of variation in the phonetic realisation of any system of phonological categories, so that more help is likely to be required from non-phonetic knowledge-sources. Secondly, whereas people are not trained to focus on the phonetic form of language (to attempt recognition in the way I have just described in outlining the PSSH), computers can be so trained, and it is an open question whether computers may not in fact prove to be far better "phonetic recognisers" than humans, given appropriate training and programming.

I conclude that the idea of attempting phonetic transcription from the acoustic record without making use of knowledge-sources above the phonetic level, is at least not an obviously foolish one.

## 1.5 The Speech Material Used in Training and Testing the Recognition System

All the speech material used in this work was for a single speaker, referred to throughout via his initials GSW. GSW is a native speaker of British English, of a non-rhotic variety, with no clear markers for either specific regional or pronounced class accent. Brought up in England with one Scottish parent, but having spent some years working in Scotland at the time of the recordings, GSW has some very slight traces of Scottish influence, which are mentioned at appropriate places in the text.

All the material used was scripted, and consisted of separate sentences of durations ranging between 3 and 6 seconds. GSW produced the utterances in a fast tempo, and in a quite informal style, so that the speech material represents many of the features of spontaneous speech, though is very largely free of hesitation, false starts, stutter, etc., and wholly free of ungrammaticality.

Two hundred of the sentences were from the SCRIBE dataset, which was designed to elicit data for as many combinations of phonemes as possible, so as to provide as much coverage as possible for context-dependent phonetic modelling. This material was supplemented with a further sixty sentences from the TIMIT dataset (Garofolo *et al.* 1990), in order to get some further coverage of phoneme combinations poorly represented by the SCRIBE dataset.

## 1.6 Brief Preview of the Content of Subsequent Chapters

In chapter 2 I list and define the phonetic categories used in the recognition technique developed in this work, and seek to show how various forms of acoustic variability can be accommodated within the phonetic framework employed. I also give a brief sketch of the system that implements the recognition technique.

Chapter 3 is concerned with cepstral representation of the short-term spectrum, and also introduces classes designed to model border-regions between phones.

Chapter 4 explains the methods used for deriving phonetic hypotheses from sequences of cepstra. The statistical basis of the transcription technique is explained in detail in this chapter.

Statistical pattern-classification techniques ideally require that healthy samples be available for each of the classes into which patterns are to be classified, and in chapter 5 I look at the problems arising from having to work with inadequate samples and at a variety of methods for overcoming such problems.

Chapter 6 is concerned with evaluation of the transcription technique, with the focus on trying to identify the contributions from individual component parts of the system, and reports the results of a number of experiments carried out to this end.

Chapter 7 presents some conclusions from the work done.

# Chapter 2

# The Phonetic Framework

## 2.1 Introduction

The main purpose of this chapter is to catalogue and define the phonetic classes used as the basis of the recognition system described in this work. A brief outline is given at the outset of statistical approaches to pattern-classification, to provide some essential context for the main discussion. Statistical approaches to speech recognition typically (though not inevitably) entail a belief in the possibility of defining segments in the acoustic record of speech that can be related to more abstract linguistic objects such as words or phonemes, and I go on to examine the reasonableness of this belief in the light of what is known about speech production. A brief overview of the recognition system, which I have called FURIDA, is then provided, as background to the account of phonetic classes that follows.[1] In giving that account, particular attention is paid to the relation between choice and definition of classes and acoustic variability arising from variations in inter-articulator timing, and there are separate discussions of assimilation and glottalisation. Throughout, an attempt is made to show that phenomena typically considered to call for a multi-tiered, feature-based approach can all in fact be handled within what is, ostensibly at least, a linear segmental framework.

[1] I have gone to the extent of christening the system simply in order to have a means of referring to it concisely. The name FURIDA is an amalgam of two Swahili words, *furaha* and *shida*, which together nicely characterise the experience of developing the system; *furaha* means "joy", "happiness", and *shida* "hardship", "difficulty".

## 2.2 Basics of the Statistical Approach to ASR

Statistical approaches to ASR, and to pattern recognition in general, are highly compatible with the account of human object-recognition argued for earlier. Speaking 'anthropomorphically', we enable the computer to recognise speech from its bare acoustic record by first providing it with numerous examples of such records together with appropriate symbolic representations of the speech (e.g. a string of phonetic or orthographic symbols) and information regarding the temporal relation between acoustic record and symbol string (which section of the record is to be associated with which symbol). The computer is thus enabled to build statistical models (cf. human mental constructs) of the kinds of acoustic measurements associated with each symbol, and can then use these models to decode further acoustic records for which no symbolic representation is given. (This sketch will be greatly elaborated in the course of this work.)

Just as in human life generally a conception founded on limited experience may prove inadequate in the face of further encounters with reality, so too in statistics it is of course a mistake to attempt generalisations on the basis of scanty evidence, and in ASR using statistical techniques it is necessary to have representative samples of the objects one wishes the computer to discern in the acoustic record, so that it can build models that will be reliable. There are, of course, a great many more words in English (perhaps 50,000 in common use) than there are sub-word units such as syllables or phonemes, and since the acoustic forms of words may be affected by the words that happen to precede and follow them in speech, it may be an onerous task to get together healthy samples of all the words one might wish to recognise in all the contexts one might wish to cover. Choice of a phonetic unit is therefore common. Such units are also affected by their context of occurrence, but we do at least start out with a significantly smaller number, and we also have the potential to cover all vocabulary-items in the language if we can build models for all phonetic contexts.

Quite simple forms of statistical modelling are used in the work described here. It is possible to model a phonetic class — one with several variants, say

— by using a mixture of Normal distributions,[2] but I have followed the general policy of trying to arrive at objects for recognition that can each be reasonably modelled using a single (multivariate) Normal distribution. In part this reflects nothing that is not simply of autobiographical interest, but it may also serve some purpose in testing the limitations of the idea that — at least for any given speaker — the acoustic record of speech may be resolved into a number of elemental forms showing merely superficial variation, forms which might prove to be relatable to terms basic to accounts of the articulatory process. However, at several points it becomes obvious that modelling with a single gaussian is inadequate, and I acknowledge these cases as they arise.

## 2.3 ASR and Linear Segmental Models of Speech

The process just described of model-building from annotated or labelled acoustic records (a process referred to as training) clearly implies a belief that the acoustic record *can* be divided into sections, and into sections that can be linked directly or indirectly (e.g. through one or more intermediate levels) with symbols which either constitute, or can be used to arrive at, a conventional linguistic representation of the utterance concerned. Just how reasonable or sensible a belief is this?

This question needs to be clearly distinguished from two questions loosely connected with it, but from which it is nevertheless quite distinct. One of these questions concerns the nature of phonological representations themselves, and in particular whether it makes sense to think in terms of *phonological* segments, of discrete units of some kind at the phonological level (whether or not these

[2]If one imagines the shape of an arbitrarily complex (non-normal) univariate distribution, one should also be able to imagine approximating that shape by choosing a number of normal curves with appropriate means and variances, and overlaying or juxtaposing them, until there is little or no difference between between the original shape and the overall envelope of the combined normal curves; in the event of a multimodal distribution, it should then be clear also that the weight given to gaussians used to model minor modes will be proportionately less than that given to gaussians modelling major modes. The probability of any particular measurement interval with respect to the original distribution may then be approximated by means of a weighted sum of its probabilities with respect to the set ('mixture') of individual gaussians.

are conceived as integral, or merely as skeletal units of some kind). The kind of *acoustic* segments under consideration here might be defended as reasonable constructs even if (as e.g. the Firthians maintain) our Phonology should actually be as dynamic and non-segmental as our (articulatory) phonetics (Local 1995), as long as the acoustic segments serve the practical purpose of providing us with a starting-point from which to recover the linguistic content of the utterance recorded.

The second question not to be confused with the one under discussion is whether it is possible to find discrete states or stages in the *articulatory* sequence responsible for the acoustic record, a question which few if any phoneticians would be happy to answer in the affirmative. The research literature universally lays emphasis upon the fact that speech is produced in dynamic fashion, with the articulators performing movements which are often overlapping and the production-system frequently showing modifications for several linguistic distinctions at one and the same time. It will in fact be one of the major contentions of this thesis that knowledge of the speech-production process — knowledge of what *can* happen in the production of particular sequences of sounds — provides the most fertile source for advances in ASR performance, but that the complexity and multi-dimensional nature of the articulatory process are capable of being accommodated from within a linear and ostensibly segmental framework, as far as decoding of the acoustic results of this process are concerned. Only one qualification needs to be made to this; while I am claiming that explicit account need not be taken of the articulatory complexity that may underlie the production of any sound or sequence of sounds, I am not claiming that higher-level structures such as syllables or feet can be entirely dispensed with in the process of phonetic classification. Indeed I shall argue explicitly further below that such structures are probably essential for phonetic classification attempted as an end in its own right, because of the need for some kind of duration-modelling at least at certain points in the transcription process.

Returning, then, to the main issue, in the case of a considerable number of sequences of two linguistic sounds, there is a more or less abrupt change in the acoustic record as one passes from the first to the second, even if the point of most abrupt change has regions to either side of it which reveal that some sort of change

was about to occur or had just occurred. In a considerable number of other cases there is no point of "catastrophe", no moment of abrupt or dramatic change, as one passes from one sound to the next (this is typically so, for example, in going from a vowel to a [w] or [r] or vice versa, or in going from a vowel to a dark lateral). In these latter cases, however, we find typical and distinctive transitions between the two sounds in question, and it is perfectly reasonable to settle upon some criterion for the placement of borders, for example placing the boundary half-way through the transition, so that the auditory colouring of "the sound on the left side of the boundary" by "the sound on the right" is minimised, and vice-versa. It is, of course, quite artificial to impose a precise and 'instantaneous' boundary in such circumstances, but it is done for a practical purpose (to enable us to gather statistics regarding the acoustic form of linguistic sounds in particular contexts) and does not reflect a belief in the 'objective reality' of acoustic segments or of isolate sounds with instantaneous beginnings and endings. In a third category of cases, it is reasonable to consider an identifiable stretch of the acoustic record as representing the result of the simultaneous production of two linguistic sounds, and in these cases labels are employed which reflect this. An example of this is the release of stops, in particular voiceless stops, when [r] is the following phone; in such cases, retroflexion of the tongue for the [r] is often underway at the point of release of the stricture for the stop, but a considerable delay may occur in the onset of voicing, and when this occurs the stretch of the acoustic record between the moment of release of the stop and the onset of voicing for the [r] is labelled with a label that makes reference to the identities of both the stop and the [r]; thus 'trb' is used for [t], [r] pairs (as in "try"), 'krb' for [k], [r] pairs (as in "cry"), and so on.

An utterance whose acoustic record has been segmented in this way thus provides a linear sequence of segments, evidently. Given that it is normal in ASR to make explicit the context of each such segment (for example by representing each one as a triphone – a phone conditioned by the two phones on either side of it), it is not possible to claim that such a linear, segmental framework *completely* ignores the dynamic nature of the underlying process responsible for the utterance. However, immediate phonetic context is by no means the whole of the story. The use of triphones clearly and explicitly takes account of the fact that the beginning

of any (artificially isolated) sound will have characteristics which result from the fact that the articulators responsible for producing it will be in process of moving toward or through the configuration required for doing so, and of moving away from the previous configuration; and similarly for the final phase of any sound. Yet the nature of the speech production process means that triphones alone are inadequate for accommodating all the kinds of variation one may find.

There is often a wide range of variation in the relative *timing* of movements of the individual articulators involved in producing a particular configuration, particularly in cases involving changes in voicing, nasality, glottal stricture or lip-rounding. For example, in laterals following [s], the initial phase of the lateral is normally voiceless, and there is considerable variation in the time it takes for voice to be restored, marked at one extreme by cases where the lateral occlusion has been released and the configuration for a following vowel is already being formed by the time phonation gets underway, with the result that the entire lateral is represented by voiceless lateral frication. Examples will be presented in section 2.5.5 below.

A major point of focus in the remainder of this chapter will be to show how phenomena such as these can be handled within a linear, segmental framework, through a combination of flexible labelling (the import of 'flexible' here will become clear as we go along), the giving of context-sensitivity to the units created, automatic expansion of basic context-sensitive phones to form subphonic elements with a largely transparent relationship to articulatory events underlying the acoustic record, and degrees of freedom in the sequencing of sub-phonic elements associated with articulatory sequences, with residual variability accommodated through statistical modelling of the spectral distributions for each subphonic element. In the next section I give a brief overview of the recognition-system as a whole, to provide the necessary context for the sections which follow, in which I define the objects used as the basis of the recognition system described in this thesis, with particular reference to the issues just discussed. This definition will also serve as an account of the principles by which I segmented and labelled training-data.

## 2.4 Brief System Overview

Between any two non-identical sounds produced in sequence, we can expect to see the acoustic results of articulatory changes of varying complexity, but given the imposition of instantaneous boundaries, in the acoustic record we count the onset of a phone to begin with the appearance of some dominant and distinctive feature associated with phones of its type (such as frication, in the case of an [s] following a vowel, or silence, in the case of a stop following a vowel), and expect only some *modification* of that feature as a result of ongoing change in the underlying articulatory process; and similar comments apply in the case of offsets of phones. In this work, training-data is labelled at phone-level for the great part, abstracting from sub-phonic distinctions of these sorts, but one of the first steps in subsequent processing is (in training-mode) to assign spectra associated with each labelled phone to just such subphonic elements, that is to onset-subphone, to offset-subphone, and where appropriate to core- or steady-state subphone. How this is effected will be described in Chapter 4. In the majority of cases, subphonic elements (subphones, henceforth) are defined in context-sensitive terms in order to indicate the source of the modification. It will perhaps be helpful to introduce the notation used for subphones at this point.

The notation used for representing context-sensitive subphones in this work is actually very simple, but demands some concentration in the first instance. Taking the word "phonetics" as an example, the acoustic record of an utterance of this word would typically be labelled as [f ax n e tc tb i kc kb s], with the 'tc' symbol denoting the closure phase of the [t], and the 'tb' symbol denoting the burst phase of the [t], and analogously for the 'kc' and 'kb' labels. This basic manual labelling would later be automatically converted to the following (or something like it):

[ SIL SIL_fB f_axA f_axB Cax ax_nA ax_nB n_eA n_eB Ce e_tcA tc tb1_iA tb2_iA tb_iB Ci i_kcA kc kb_sA kb_sB s_silA sil ].

Two basic categories of label can be distinguished in this example. On the one hand are the labels which contain an underscore and end in either an 'A' or a 'B'; and on the other are those labels which do not. Labels in the second category

are not marked for context. In the example, the labels falling under this second category are [SIL] (initial silence), [Cax] (representing the core or steady-state subphone of a schwa), [Ce] (representing the core or steady-state subphone of [e], the vowel in "Ted"), [tc] (the closure for the [t]), [Ci] (representing the core or steady-state subphone of [i], the vowel in "sit"), [kc] (the closure phase of the [k]), and [sil] (final silence). (The justification for treating such cases as context-independent will be presented in due course). Turning to the symbols in the first category, those with underscores and terminating 'A' or 'B', the first thing to be said is that all these labels denote parts of phones, and parts of phones in particular phonetic contexts. In symbols ending in 'B', the element preceding the underscore indicates the preceding context, that is, the phone most recently produced, while that following the underscore represents the onset of the current phone; thus 'SIL_fB' denotes the onset subphone of an [f] articulated following an initial silence, 'f_axB' denotes the onset subphone of an [ax] following an [f], and so on. In strings ending in 'A', by contrast, the element following the underscore, minus the terminal 'A', represents the phone produced next, that is, the following context, while the element preceding the underscore represents the offset of the current phone; thus 'ax_nA' denotes the offset subphone of an [ax] prior to the articulation of an [n], 'n_eA' denotes the offset subphone of an [n] prior to the articulation of an [e], and so on.

Figure 2.1 may prove helpful in fixing in the reader's mind the distinction between 'A'-terminating and 'B'-terminating subphone labels. It may also be useful to associate the 'B' with the phrase "Beyond the phone-boundary", and the 'A' with the phrase "in Advance of the phone-boundary". To further strengthen the reader's grasp of the notation, two more examples follow. Given a vowel [ii] (as in "bean"), occurring between an [b] and an [n], the automatic expansion of the vowel symbol would produce three subphone symbols, namely

[ bb_iiB Cii ii_nA ].

Given a [z] (as in "gazelle"), occurring between a schwa ([ax]) and a vowel [e], the automatic expansion of the [z] would produce two subphone symbols as follows:

[ ax_zB z_eA ].

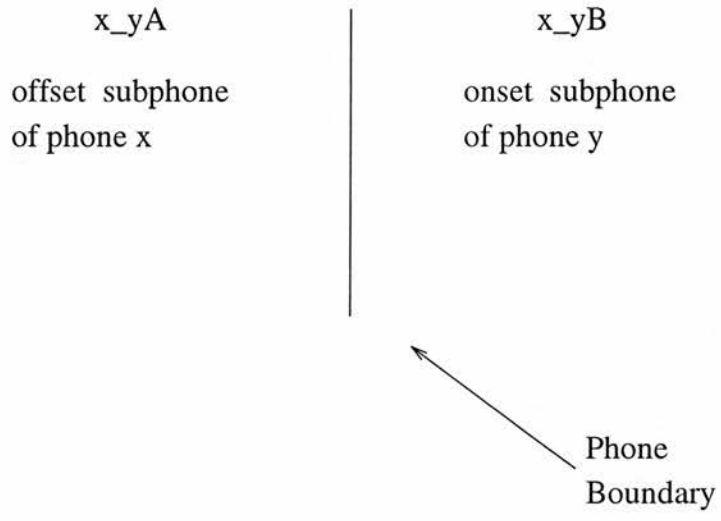



**Figure 2.1.** Transcription of Subphones at Phone Border Between [x] and [y]

Further assistance with the subphone symbols will be offered to the reader at each point where such symbols are used in this chapter, either in the main text or in footnotes. All the symbols used for specific phones are brought together in Appendix A.

Once the training-data has been allocated to subphonic classes, statistical parameters are estimated for each class and subsequently used in recognition-mode to perform probabilistic subphonic classification of unlabelled acoustic records of speech. Recognition begins with calculation of class-membership probabilities for each sliver of the acoustic record of the utterance to be recognised, resulting in a matrix of probability-scores, and the best subphonic transcription is found as the most probable path from the first to the last column of the matrix, possible paths being highly constrained owing to the context-sensitive definition of the subphonic classes (an [ii_sA] assignment, for example, can be followed only by another identical assignment or by an assignment to [ii_sB]).[3] This initial output will be referred to below as the *AS-transcription* (automatic subphonic

[3]The [ii_sA] subphone is the offset subphone of an [ii] vowel occurring before an [s], and the [ii_sB] subphone is the onset subphone of an [s] following an [ii]; clearly then, a classification of a sliver of the acoustic record as an [ii_sA] is consistent only with the classification of the next such sliver as either another [ii_sA] or as an [ii_sB].

transcription). It is subsequently converted automatically to a transcription that uses just those classes (more or less) used in manual labelling of training-data, so as to make possible a comparison of automatic and manual transcriptions where both are available; this converted transcription is referred to below as the *AP-transcription* (automatic phonetic transcription).

It should be noted that while the context-sensitivity conferred on subphonic elements is initially quite specific (so that, for example, the onset of an [r] following an [o] (as in "borrow") gets converted to an [o_rB]), there is usually not sufficient data for statistical modelling of completely specific subphone-classes. Procedures are therefore invoked (as will be described in Chapter 5) for merging where necessary those subphones likely to have similar acoustic profiles (such as, continuing with the example, [o_rB] (as in "borrow"), [uh_rB] (as in "furrow") and [aa_rB] (as in "Cara"), which might combine to form a generalised subphonic class denoting onset of [r] following any of these three vowels).

The decision not to use triphones was motivated chiefly by the relative lack of training-data, though it also seemed — after lengthy consideration of spectrographic data — that for the majority of cases variation that could be accounted for in terms of immediate phonetic context at all could be reasonably accommodated piecewise (with the first phase of a phone affected primarily by left- and the final phase by right-context). In a small number of cases where this was manifestly not the case — as for example with laterals — allophones were first distinguished, and the allophones then subjected to subphonic analysis as with other phones. Further details regarding such cases will be given shortly.

## 2.5 Segmental Units, Sequence-Constraints and Timing-Variation

The basic inventory of phonetic categories used in FURIDA reflects the author's understanding of the phonological system used by GSW – basically, a standard form of Southern British English (with one or two peculiarities that will be discussed further below). A number of allophonic categories are employed in addition, to make the Normality assumption more reasonable for the categories

concerned. Subphonic analyses – resolving phones into smaller sections – were arrived at in the light of a number of considerations. Firstly, a subphonic resolution was required to lead to a situation where the Normality assumption would be more justified than would have been the case for a whole-phone representation (whether or not allophonic distinctions had been introduced for the category in question). Secondly, the fineness of a subphonic analysis was limited by the need to have sufficient data to allow statistical modelling of the subphones decided upon. It might, for example, have turned out that advantages could be got by modelling specific ordinal positions within the temporal development of individual phones (conceiving phones as ordered sequences of spectra), but such a course of action would not have been practical with the amount of data available, quite apart from any other difficulties it might have entailed. Thirdly, the subphones had to correspond in a more or less obvious way with phonetically interpretable parts of phones. The overriding aim of the work, after all, was to produce a system for transcription that could be understood in phonetic terms. Finally, the overall scheme must be as simple as was consistent with reality, where reality included both the phonetic facts and the amounts of data actually available for training. Not least, this was of some significance for the writing of programs, where, for example, having to treat different categories in a very large number of different ways would have made the programming task extremely tedious.

No short and simple formula is available for summarising the variety of subphonic analyses to be described in the rest of this section, but the following picture of what the author sees as the most general features of the acoustic (spectrographic) data may be of some use. Within the confines of phones (accepting all the qualifications made earlier about the artificiality of instantaneous boundaries and isolated sounds) we find in some places patterns that are sustained, that exhibit no consistent change over some tens of milliseconds, while in others we see periods of change reflecting underlying articulatory changes. In some cases, the salience of change may be such as to dominate our perception, while in others (in some fricatives for example), the change may be so slight as to be outweighed by the perception of sameness. Generally speaking, some phones lent themselves to a subphonic analysis that might be expressed in terms of "change toward a pattern, maintenance of the pattern, change away from the pattern", others to

| Fricative | Example |
|---|---|
| [s] | "so" |
| [f] | "fee" |
| [th] | "thumb" |
| [sh] | "shoe" |
| [z] | "zoo" |
| [v] | "vie" |
| [dh] | "the" |
| [zh] | "genre" |

**Table 2.1.** Fricative Labels

an analysis rather as "change toward a pattern, change away from the pattern", and others to an analysis as "maintenance of a single dominant pattern but with evidence of ongoing underlying articulatory changes". In the third of these three cases it seemed reasonable to suppose that the earlier portion of the phone would be affected by the earlier phase of the process of underlying change, and the later portion by the later phase (where most activity would be part of the process of getting ready for the next sound to be produced); even in these cases, then, where the acoustic changes were 'secondary', an analysis in terms of onset and offset seemed reasonable. Hence just two basic patterns of subphonic analysis are used repeatedly for a large number of the phonetic categories involved: that of onset, steady-state, offset, and that of onset, offset. In some cases – for example where glottalisations occur – other patterns may be overlaid on one or other of these basic patterns, and in the case of some categories of phone, such as pre-vocalic aspirated stops, rather different analyses were employed. The sections that follow tell the story in its full detail.

### 2.5.1 Fricatives

The basic phone-level symbols used for fricatives are listed in table 2.1. With [-VOICE] fricatives no allophones are distinguished except in the case of [th] (the first sound of "think"), where the stricture-gesture is sometimes excessive for the production of a fricative, leading to a complete blockage of the oral tract and

so to a stop-like acoustic realisation; in these cases, instead of the usual single [th] label, a pair of segments are used, [thS] and [thR], one for the closure-phase and one for the release-phase. The 'thS' label is used for the region of silence corresponding to the closure phase, ('S' is mnemonic for *stricture*), and the 'thR' label is used for the region of frication that follows ('R' is mnemonic for *release*). (Any cases involving silence that do not fit this simple closure-release account, for example that begin with frication, have a period of silence in the middle, and end with further frication, are labelled simply as [th], the main reason for this being that examples were not sufficiently numerous to allow modelling of a distinct class or classes). With the exception of this case of [thS] and [thR], the division into two subphonic moments — the onset left-context-sensitive and the offset right-context-sensitive — is effected automatically for all fricatives.

The two-subphone treatment of [-VOICE] fricatives is actually too crude, and in the case of [s] in particular leaves the door open for segmentation errors, such as the insertion of a [t] in the latter stages of the [s].[4] I now briefly consider some of the causes of this, and possible solutions. In the change from [s] to a following vowel, the glottis must change from a wide open to a closed position, the constriction made by the front of the tongue must be relaxed, and the tongue-body must be put into the correct position for the vowel. Small variations in the relative timing of these gestures, and in their force, may have a major impact on the acoustic form of the final stages of the [s]. If the glottal closing gesture is relatively late, with the alveolar constriction already approaching the point where frication ceases to be viable before the airflow has ceased, air will tend to rush out and give rise to aspirative formants revealing the extent to which the tongue-body has already been put into position for the following vowel (figures 2.2 and 2.3).[5] A weak transient frequently appears just before these formants do,

[4]By insertion is meant the 'recognition' by the system of a phone which is not actually present, a system hallucination so to speak; in the example given in the text, the system thinks that the final acoustic material for what is actually an [s] provides evidence for postulating the existence of a [t], (while typically recognising the earlier part of the acoustic material for [s] as an [s].

[5]In all the spectrograms shown in this work, 1 kHz intervals are shown by crosses in the vertical dimension, covering the range from zero to 8000 Hz, while tics along the horizontal axis mark intervals of 10 milliseconds. The frequency scale is shown explicitly for the first spectrogram.

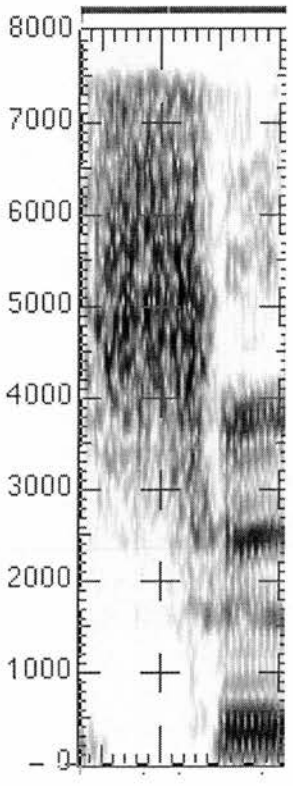



**Figure 2.2.** [s ir] from utterance of "considered" (sc003)

or some time thereafter (figures 2.4 and 2.5). The degree to which the constriction has been relaxed, in cases of relatively late glottal closure, reflects itself in the acoustic record: a more relaxed oral stricture may see frication persist, but with a qualitative change, while after a certain point no frication at all may be evident, with *only* aspirative formants testifying to the fact that the glottis has still not reached the degree of closure required for voice (figure 2.6). If the glottal closing gesture is relatively early, on the other hand, we will typically find a short silence-interval between the [s] and the vowel, with interruption of the airflow obviously making it impossible to sustain frication (figures 2.7 and 2.8).

The acoustic variation caused by these variations in timing is poorly handled by a simple onset-offset modelling of prevocalic [s]'s, at least when single gaussians are used for the modelling, and something more sophisticated is required. One could stay with the onset-offset analysis, and model the offset with a mixture of

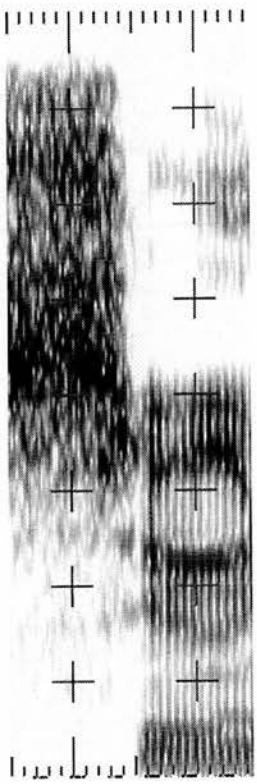


**Figure 2.3.** [s uuf] from utterance of "soon" (sc059)

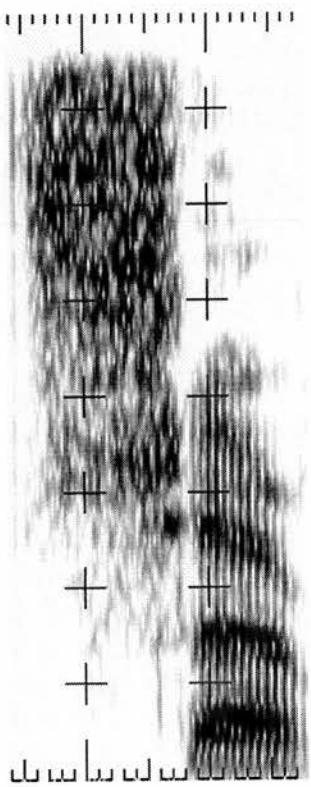


**Figure 2.4.** [s e] from utterance of "several" (sc018)

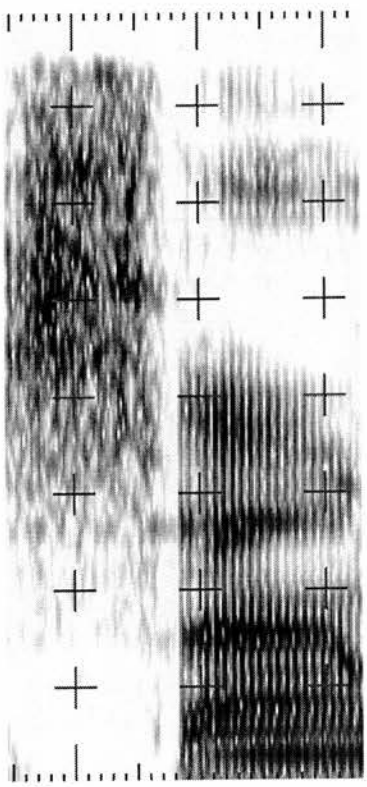


**Figure 2.5.** [s a] from utterance of "sandwiches" (sc052)

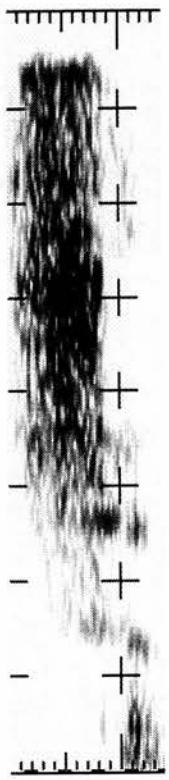


**Figure 2.6.** [s ax] from utterance of "habits of" (sc008)

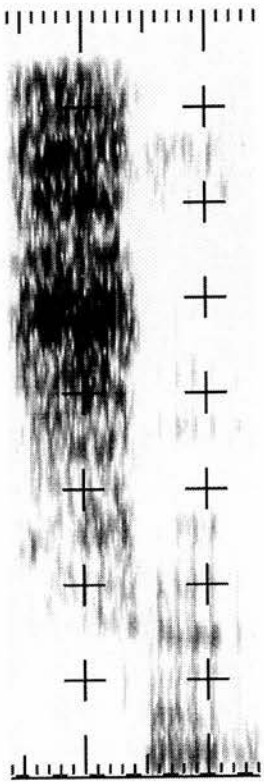


**Figure 2.7.** [s ax] from utterance of "since" (sc017)

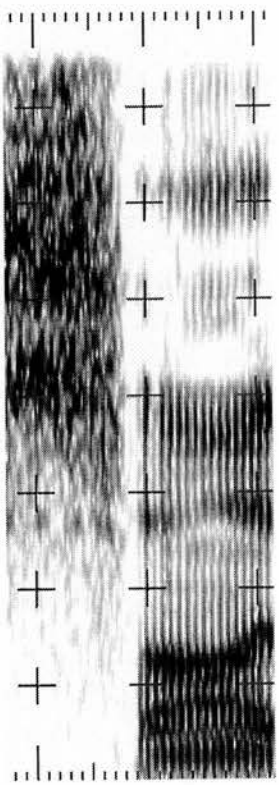


**Figure 2.8.** [s ai] from utterance of "decidedly" (sc105)

gaussians rather than a single one; alternatively, one could use a larger number of subphones, with perhaps obvious sequencing-optionalities for the additional ones, retaining the simple modelling assumption; a third subphone might be used to cater for tokens with late glottal closure and distinct pre-formants, and a fourth to model silence-intervals preceding the vowel. The reader may perhaps be feeling that this is all very slight, and perhaps even that it is a lot of fuss about nothing, but the greatest possible accuracy and sensitivity is required if mistakes are to be avoided or minimised, and I present two spectrograms that illustrate the point, in figure 2.9. One of these is for the [s] in "say", and the other represents the [s T] in "instead",[6] both being taken from the utterance, "My mother gets cross when I say "yeah" instead of "yes" ", and the figure demonstrates the attention to detail that is required for any serious attempt at phonetic recognition. The [T] of "instead" is merely hinted at, so to speak, in the speaker's production of the utterance, but in the writer's judgement is certainly present. Accurate discrimination of [s] from [s T] clearly requires very close and detailed modelling if it is to be achieved at this level (that is, by reference to the acoustic record alone).

It is rare to find voicing for a following sound starting during a [-VOICE] fricative, but common to find a (usually quite slight) delay in the offset of voicing from a preceding sound at the beginning of one of these fricatives. I have left this as a form of residual variability to be accommodated via the statistical modelling, partly through lack of data once context-sensitivity has partitioned the data, partly because there is not always a great deal to choose between early context-specific spectra from a single context whether there is any carryover voicing or not, and partly because the use of border-straddling classes, to be explained in Chapter 3, reduces the scope of the problem slightly. I shall return to this problem shortly in the course of discussing voicing-assimilation in [+VOICE] fricatives.

[+VOICE] fricatives present rather more difficulty than their [-VOICE] counterparts, just because of the presence of voicing as a component of their acoustic realisation. No allophones are distinguished for [z] (the first sound of "zebra") or for [zh] (the first sound of "genre") or for [v] (the first sound of "very"), but

[6] I use 'T' to denote the unaspirated form of [t], and abstract for the purposes of the example from the closure-release structure of the stop.

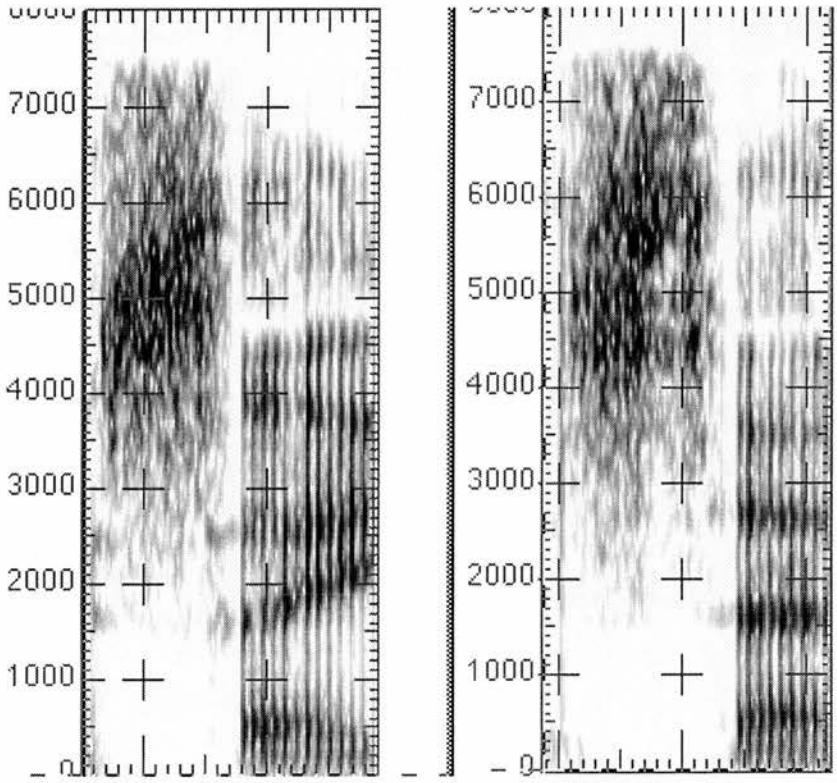



**Figure 2.9.** [s] (on left) and [s T] from utterance sc104 (see text) illustrating the difficulty of [s] vs [s T] discrimination

a number of allophones are distinguished for [dh] (the first sound of "then") in order to try to reduce the very extensive range of variability in its realisations.

Most [dh]'s are distinguished in labelling solely in accordance with the VOICE feature of the phones preceding them, giving us post-[+VOICE] [dh], labelled 'PVdh', and post-[-VOICE] [dh], labelled 'PNVdh'. Sentence-initial [dh]'s are also distinguished, and labelled 'INITdh', because of their frequent stop-like realisation. Post-nasal [dh]'s are labelled 'PNASdh' regardless of the specific identity of the nasal; these phones typically have a very short duration of about 10–25 ms. Both [INITdh]'s and [PNASdh]'s clearly have a degree of left-sensitivity built in to their names, and so are treated subsequently as single-subphone phones sensitive only to their right contexts, giving us, for example, [INITdh_eA] at the beginning of sentence-initial "Then ...", and [... PNASdh_eA ...] as part of the subphonic transcription of "and then ..." (with the [d] of "and" deleted). As with [th], the approach for [dh] may be executed with a measure of force that results in a total obstruction of the vocal tract, and so to a period of silence or merely residual voice; if the stricture is then relaxed before progression to the next sound, the acoustic result may be more reminiscent of a stop-release. The 'allophonic' labels 'dhS' and 'dhR', parallelling 'thS' and 'thR' detailed above, are used in these circumstances. [dhS] subphones are treated as context-independent, and [dhR] subphones as sensitive only to their right contexts.

[v]'s show a great deal of variability, and in a number of different dimensions, and an attempt could perhaps have been made to allocate them — before conferring context-sensitivity — to one of a number of allophonic groups at the stage of manual labelling. Thus tokens coming closest to approximant-like realisations (distinct pitch-pulses visible across the frequency-range, a clear formant-pattern, and little or no frication) could be labelled as 'vApp', those coming closest to fricative-type realisations, with a temporal development not indicative of a closure-release articulation, could be labelled as 'vFr'; while tokens with a clear closure-release structure, with the closure phase either silent or with only residual voice, could be labelled as 'vSR', or as 'vS' in the event of there being no release. This 'allophonic' division would not have been very satisfactory, since the variability involved defies any simple categorisation. Without some such division,

however, it is difficult to recognise certain tokens of [v] correctly even after the introduction of context-sensitivity and subphonic analysis, with [vSR]-type tokens, for example, typically scoring more highly as [b]'s. [v] represents a sound for which piecewise context-dependency (taking the left subphone as dependent only on the preceding context, and the right subphone as dependent only on the following context) is quite inadequate, and perhaps constitutes another case where triphone-modelling, and perhaps (even for a single speaker) mixture-modelling too, are necessary. Similar remarks may be justified in the case of all [+VOICE] fricatives, the cue to the inadequacy of piecewise modelling being the importance of the voicing environment in which the [+VOICE] fricative occurs, where this includes the VOICE feature of the phones on both sides of the fricative.

The immediate voicing environment appears to be one of the most significant influences on the acoustic realisation of [+VOICE] fricatives in general. Two factors are involved in this: firstly there is the relatively great inertia involved in the system for producing voice (relative to the systems for producing changes in things like tongue-tip position, for example), and the other is the relatively great effort required for production of simultaneous voice and frication, with the oral stricture leading to an increase in mouth-pressure and so to a drop in the transglottal pressure-difference required for voice. Voicing can be prolonged by providing a boost to subglottal pressure through an exertion of the expiratory muscles, but the relatively massive inertia of the latter makes them unfavoured as an instrument of short-term segmental differentiation (Ohala 1989). No doubt this is at least part of the reason why the +VOICE/-VOICE distinction in respect of word-final English fricatives is effected in large measure by lengthening of the preceding vowel or sonorant.

In continuous speech we normally find a significant measure of voicelessness in [+VOICE] fricatives preceding [-VOICE] sounds. It does not follow from the fact that a [+VOICE] fricative is wholly voiceless, that all supraglottal changes required for passing to the fricative from the preceding sound must have been completed by the time the fricative begins, but it often appears to be the case that they are. In the system being described, the extreme in this form of variation — complete devoicing of the fricative — is handled by allowing [+VOICE] fricatives between [+VOICE] and [-VOICE] sounds to be realised by an 'offset'

subphone alone (conditioned as usual on the right context). In cases where this freedom is invoked, we act as if all supraglottal changes required to arrive at the fricative articulation are completed while the sonorant is maintained — or at least that a point has been reached where the switch to the fricative can be effected more or less instantaneously — so that as soon as frication comes to dominate the spectrum, any subsequent modification will be the result of adjustments beginning for the production of the *next* phone (the one that follows the fricative). This may not be a perfectly accurate representation of what happens in these circumstances, but it seems to be a passable approximation to it. Not mixing data from cases having initial voicing with cases devoiced throughout means that the Normality assumption is not grossly violated, while, clearly, allowing the immediate passage from vowel or nasal or lateral to an offset subphone of the [+VOICE] fricative means that improbable matches of data to the model for the onset subphone for the fricative are not forced on the system.

Devoicing of [+VOICE] fricatives is usually a good cue to the VOICE feature of following sounds. Generally speaking a markedly devoiced [+VOICE] fricative signals a [-VOICE] sound to follow, although devoicing may just arise because of fading voice-effort as at a clause-boundary, even where the sound that follows across the clause-boundary is also [+VOICE]. (The annotation of phones for higher-level features such as position with respect to phrase- or clause-boundaries would be possible, (and indeed highly desirable, if modelling with single gaussians), if working with huge amounts of training-data, but was not really practicable in the present context.)

Effects of the place of articulation of following phones on final subphones of fricatives are very reliable (and also uniform over seven of eight speakers whose speech I have examined in detail). The distinctiveness of the alveolar context vis-a-vis both the labial and velar is noteworthy, and fails only when the sound following the alveolar phone is [r], because of anticipatory lip-rounding during the fricative. Examples and further discussion will be found in the next section.

Table 2.2 summarises the details of allophonic symbols used for fricatives or fricative subphones. Table 2.3 summarises the context-dependencies of the various fricative subphones.

| Class | Explanation |
|---|---|
| [thS] | stricture-phase of a [th] given stop-like realisation |
| [dhS] | stricture-phase of a [dh] given stop-like realisation |
| [thR] | release-phase of a [th] given stop-like realisation |
| [dhR] | release-phase of a [dh] given stop-like realisation |
| [PVdh] | [dh] following a [+VOICE] phone |
| [PNVdh] | [dh] following a [-VOICE] phone |
| [PNASdh] | [dh] following a nasal |
| [INITdh] | sentence-initial [dh] |

**Table 2.2.** Fricative Allophones

| | |
|---|---|
| [s],[f],[th],[sh] | piecewise left-/right-context-dependent (i.e. left subphone left-context-dependent and right subphone right-context-dependent) |
| [z],[v],[zh],[PVdh] | piecewise left-/right-context-dependent when the following phone is [+VOICE]; when the following phone is [-VOICE] an alternative subphonic expansion is possible representing the fricative as a right-context-dependent 'offset' subphone |
| [PNVdh] | piecewise left-/right-context-dependent |
| [thS],[dhS] | context-independent |
| [thR],[dhR] | right-context-dependent |
| [PNASdh] | right-context-dependent |
| [INITdh] | right-context-dependent |

**Table 2.3.** Context-Dependency of Fricative Phones and Subphones

| Stop | Example |
|---|---|
| [t] | "Tom" |
| [p] | "Pam" |
| [k] | "Kim" |
| [d] | "Don" |
| [b] | "Ben" |
| [g] | "Garry" |
| [T] | "sty" |
| [P] | "spy" |
| [K] | "sky" |

**Table 2.4.** Stop Labels

### 2.5.2 Stops

Abstracting for a moment from closure-release structures, I introduce the discussion of stops with the help of table 2.4. This is merely for the sake of exposition, the labels in this table not actually being used in the labelling of data (the labels used in labelling, and in subsequent automatic expansion, are however derived from the labels in the table, and are themselves tabled and discussed further below).

Closure- and release-phases are labelled separately (thus [tc] [tb], [gc] [gb], etc.) where they occur. The closures of unaspirated stops [T], [P] and [K] are labelled as [utc], [upc] and [ukc] respectively, where the 'u' is meant to be mnemonic for 'unaspirated'. Where two stops follow in sequence and the first is unreleased, if there is no obvious acoustic mark of a switch from one to the other a complex closure-label is used (thus [tdc], [bgc], etc.). The symbol 'tdc', then, represents the closure of an unreleased [t] combined with the closure of a following [d]. Closures are one of a small group of segments that are given no context-sensitivity, and although the identity of the stops they belong to is encoded in the manual labelling, the lack of acoustic distinctiveness in the closures themselves leads to their being automatically converted to 'generalised' forms later on.[7] Closure

[7]The initial labelling was designed to be as information-rich as possible, since at the outset it was impossible to predict all the uses that one might have wished to put the data to; labelling

| VOICE of Preceding Phone | Manual stop label | Automatic conversion to |
|---|---|---|
| - | [tc] | NVSAc |
| + | [ttc] | NVSAc |
| + | [bc] | VSAcv |
| - | [bc] | VSAcnv |
| + | [dbc] | VSAcv |
| + | [tdc] | NVVSAc |
| + | [dtc] | VNVSAcv |

**Table 2.5.** Conversion of Specific Stop Closure Labels to Generic Labels

labels are converted according to their VOICE features, with an additional 'diacritic' ('nv' or 'v') in certain cases to indicate the VOICE feature of the phone preceding them (this having some bearing on the likelihood of any 'voice-bar'), as the examples in table 2.5 show. The generic labels all contain 'SA', which is meant to be mnemonic for "stop or affricate", the closures for affricates being generally treated along the same lines as those for stops. The generic labels also begin either with 'V', 'NV', 'NVV' or 'VNV'. 'V' is used for the closure of a [+VOICE] stop or pair of stops or a stop affricate pair, 'NV' for the closure of a [-VOICE] stop or pair of stops or stop affricate pair, 'NVV' for the closure of a pair of stops, or a stop afficate pair, when the first member of the pair is [-VOICE] and the second is [+VOICE], and 'VNV' for a similar situation but with the VOICE features reversed.

Note that where both of a pair of consecutive stops have the same VOICE feature, the generic label does not explicitly indicate that a pair of stops is involved; thus a 'NVSAc' label may designate either a single [-VOICE] stop closure or a pair of such, and a 'VSAcv' or 'VSAcnv' label may designate either a single [+VOICE] stop closure or a pair of such. Note also that the 'n' and 'nv' diacritics are used only when the first stop (in the event of single stops, the only stop) is [+VOICE].

Unaspirated stops are explicitly marked as such when they are prevocalic

with generalised labels from the bginning would have meant the throwing away of information that might have proved useful in the long term.

(there simply was not sufficient data to model them explicitly everywhere) (thus [utc] [Tb], [upc] [Pb], [ukc] [Kb], in distinction to aspirated [tc] [tb] etc.). In complex-closure labels, however, the aspirated-unaspirated distinction is collapsed as far as the closure is concerned, so that there are no labels like 'putc' (the symbol 'putc', had it been used, would have denoted the closure of an unreleased [p] combined with the closure of a [T], as might occur in a production of the word "apt"). Flapped [t]'s are labelled 'tflap', and lenited [+VOICE] stops (with weak closing gesture and consequently no significant pressure build-up to feed a definite release) are labelled 'dnc', 'bnc' or 'gnc' as appropriate. The 'nc' here is meant to be mnemonic for "no closure", so that [dnc] is a [d] that has no fast closure (and so, obviously, no closure-release structure). The presence or absence of a definite release was the deciding issue in all cases where closures were incomplete (so that one or more higher formants remained visible within the closure); hence the distribution for any stop-closure would generally include cases with formant-bands though these would be a minority. Realisations of the -ed morpheme after phones other than [d] or [t] are counted as [d]'s when following a [+VOICE] and as [T]'s when following a [-VOICE] sound.

Aspirated pre-vocalic releases are modelled with two subphones, with the first designed to capture the phase of the release where frication is dominant, and the second the phase where aspiration is dominant. Both subphones are subsequently annotated for right context (so that an utterance of "two", for example, might be labelled as [tc tb uu] and subsequently expanded to [SIL tc tb1_uuA tb2_uuA (tb_uuB) Cuu uu_silA sil], where round brackets indicate an optional subphone). Pre-vocalic and aspirated [-VOICE] stops provide a well-known and much-studied example of inter-timing variation. There is a range of variation in the degree to which the articulatory repositioning for the vowel is complete by the time that voicing gets underway, with generally an inverse relationship between duration of the release of the stop and duration of the onset subphone of the vowel, any repositioning that has been done in the aspirative phase of the stop-release not having to be done once voicing is underway. Detailed exploitation of this relationship has not been built into the system for lack of time, but optionality of the vowel onset following these stop-releases accommodates the basic variability to some extent.

Pre-consonantal and unaspirated pre-vocalic releases are modelled with a single subphone subsequently annotated for right context. Releases of unaspirated [-VOICE] stops are typically very brief, especially when occurring as part of [s T], [s P], and [s K] clusters, but it has to be recognised that the aspirated/unaspirated distinction is reflected in a continuum of variation. (Catford 1977) states that unaspirated sounds are marked by a narrowed glottis, and that in unaspirated stops, the increase in oral pressure during the closure is already flattening off or even falling by the time the stop is released, with the vocal folds by that time almost in position for voicing, this being the reason for the prompt resumption of voice and the absence of an aspirative interval. In labelling it is always a blessing to be able to use purely structural criteria if attempting allophonic discrimination (e.g. to let the allophonic label be determined by the broad class of the preceding or following phone, so that no agonising is required over what allophone is present), and a curse to have to categorise on the basis of spectral detail, and certainly most of the stops labelled as unaspirates in this work were part of [s stop] clusters (if 'cluster' be allowed a loose interpretation), or realisations of the past tense marker for regular verbs with stems ending in [-VOICE] sounds other than [t], or in second position in [k t] or [p t] clusters (thus "act" as [a ktc Tb], "slept" as [s l e ptc Tb]), or in second position in [f t] clusters (thus "after" as [ a f utc Tb ax ]). But there were a few cases where stops were labelled as unaspirates even though they did not occur in the standard structural positions, though in all cases these were stops occurring after [s], [k] or [f] across a word-boundary, preceding an unstressed vowel. (The fact that such stops could conform more closely to the unaspirated than to the aspirated archetypes may suggest that there is a physiological basis to the distribution of unaspirated stops, rather than a phonological one, but that is not strictly an issue here.)

The context-dependencies of the various stop phones and subphones are summarised in table 2.6.

### 2.5.3 Affricates

Closures are processed in the same way as stop-closures, while frication phases are treated in the same way as fricatives. The [-VOICE] affricate (as in "chew")

| | |
|---|---|
| [tc],[pc],[kc] | context-independent |
| [utc],[upc],[ukc] | context-independent |
| [dc],[bc],[gc] | context-independent |
| [tb],[pb],[kb] | right-context dependent (2 subphones when prevocalic) |
| [Tb],[Pb],[Kb] | right-context-dependent |
| [db],[bb],[gb] | right-context-dependent |
| [dnc],[bnc],[gnc] | piecewise left-/right-context-dependent |
| [tflap] | piecewise left-/right-context-dependent |

**Table 2.6.** Context-Dependency of Stop Phones and Subphones

| Nasal | Example |
|---|---|
| [n] | "no" |
| [m] | "my" |
| [ng] | "thing" |
| [nsyl] | "hidden" |
| [msyl] | "Mmm!" |
| [ngsyl] | "engrossed" (fast speech) |

**Table 2.7.** Nasal Labels

may be represented by 'ch' when abstracting from internal structure, and the [+VOICE] affricate (as in "Joe") similarly by 'jh'. The labels 'chc' and 'chb' are used in manual labelling for the closure and frication phases respectively of [ch], and similarly 'jhc' and 'jhb' are used for the closure and frication phases respectively of [jh]. The use of generalised labels for combined closures of unreleased stop and following affricate has already been covered in the section preceding this one.

### 2.5.4 Nasals

The basic phone-level labels for nasals are shown in table 2.7. In general (for the speaker GSW), nasals are more often than not quite insensitive to the F2 frequencies of neighbouring vowels, and are strongly sensitive to the VOICE environment in which they occur, with the whole murmur typically being weak

early in a nasal that follows a [-VOICE] sound, and likewise late in a nasal that precedes such a sound. But prosodic factors interfere with this latter tendency, and one may find very robust nasal murmurs before [-VOICE] sounds in certain conditions. I now discuss each of these points in more detail.

[n] and [ng], no less than any other consonants, are subject to contextual influences in respect of the precise placing of the tongue for the oral closure (consider, e.g. the well known tendency to dentalise [n] before a consonant such as [t] or [d]), each having an acceptable area for the stricture rather than a precise point, so that the speaker has some leeway in producing one of these sounds between any two phones, the precise location of the closure probably being related to whatever happens to be the easiest articulatory route to follow in passing from the preceding sound through the nasal to the following sound. However, while we find occasional cases where the formant pattern in one of these nasals is clearly responsive to the frequency of F2 in neighbouring vowels (figures 2.12, 2.13 and 2.14), in many other cases the formant pattern in nasals seems quite insensitive to the frequency of F2 in neighbouring vowels, as exemplified in figures 2.10 and 2.11. Triphone-modelling might be superior to piecemeal left-right modelling in the former cases, but given the large numbers of cases of the latter kind it may be an extravagant use of resources. Even left-right modelling with piecemeal context-sensitivity often appears to be an extravagance in the case of intervocalic nasals, where very little change may be evident in the formant pattern throughout the nasal. There may well be a case for having a single model for all intervocalic [n], and similarly for [m] and [ng], and modelling with a mixture gaussian to accommodate variation in the frequency of F2 whether in nasals with level F2 or those with clear influence from F2 in neighbouring vowels. In FURIDA, left-right context-sensitive modelling was used for nasals as for most consonants, but I have to acknowledge that breaking up the data according to the species of the left vowel, and according to the species of the right vowel, often does not make a lot of sense.

The VOICE environment can have a quite major impact on the form of nasals. The whole nasal spectrum is sensitive to such factors, and the second formant particularly so. Thus a nasal will tend to gather strength, with change usually most evident in F2, in a nasal that begins a word or an utterance, or that follows

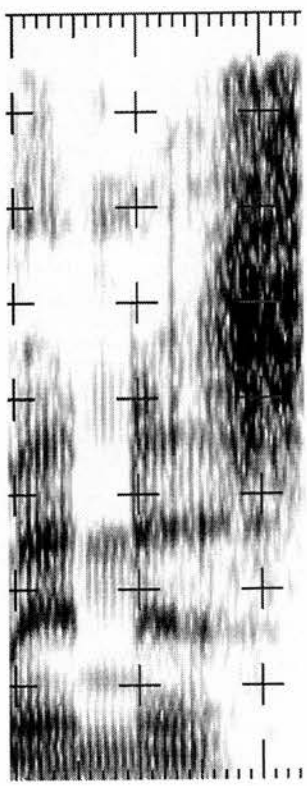


**Figure 2.10.** [i n ir s] from utterance of "minister" (sc008)

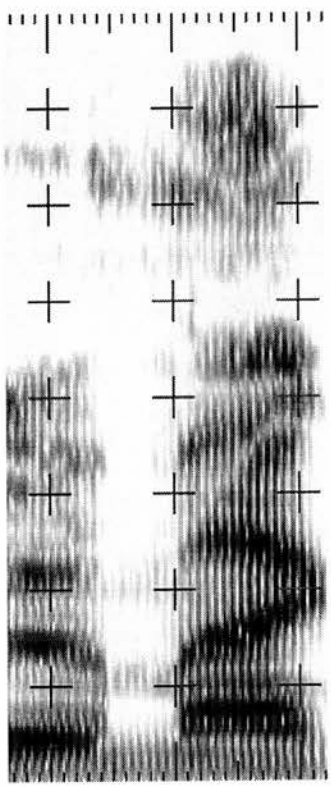


**Figure 2.11.** [ax m a] from utterance of "a magazine" (sc020)

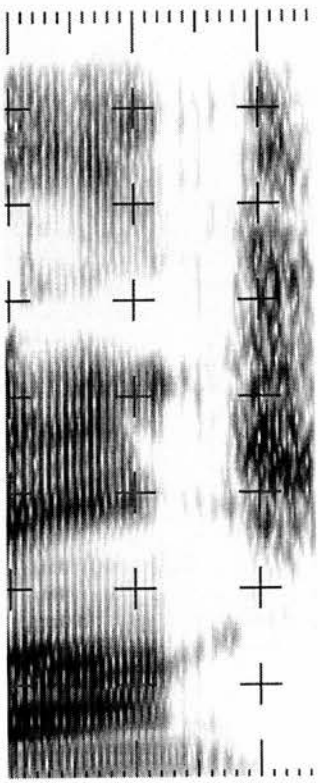


**Figure 2.12.** [aa n] from utterance of "Blanche" (sc107)

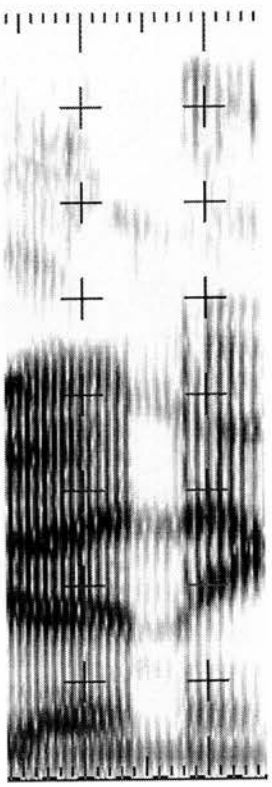


**Figure 2.13.** [e n i] from utterance of "Jennings" (sc107)

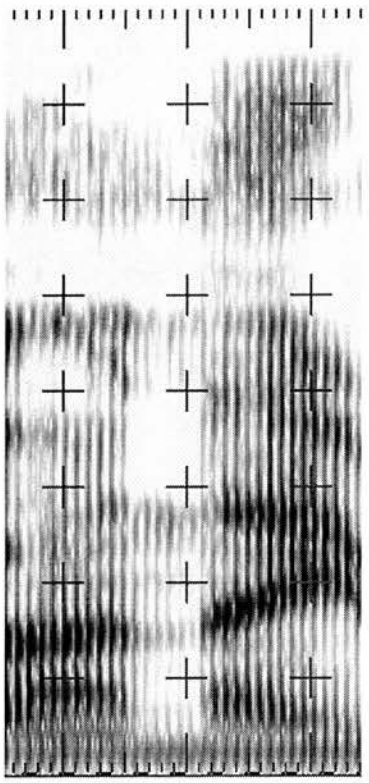


**Figure 2.14.** [aD1 n e2ii] from utterance of "our neighbours" (sc118) (where [aD1] represents a realisation of the diphthong [au] as in "now"

a voiceless sound that ends a preceding word or syllable, and to weaken in a nasal that precedes a voiceless sound in the same syllable. Frequently the gain in or loss of energy is more significant than any movement in nasal formants associated with the place of articulation (POA) of the preceding and following sounds, and left-right modelling and context-sensitivity is thus more relevant through its highlighting of the former effects, in cases where a switch in VOICE occurs.

It would be simplistic to put all the emphasis on the VOICE environment, however, important though this this, since prosodic factors also come into play. This is illustrated in utterance 114 in the phrase "his own photographs", where we have an [n f] sequence in which the nasal is robust throughout, in spite of the [-VOICE] feature of the [f], presumably because of the prominence the speaker gives to the word "own" (figure 2.15); the pre-[f] [n] in utterance 127 in the phrase "before we can finish the garment", on the other hand, is very weak in its offset, presumably because the function-word "can" is very lightly articulated (figure 2.16). Not that a simple distinction between function-words and content-words is adequate to explaining the phenomena. In utterance 146, for example, in the phrase "has been served", we find a robust and sustained [n] in an unstressed function word (figure 2.17). (Looking at such cases, one is tempted to suggest that nasal articulations, through presenting relatively great supraglottal resistance, may be sustained by a speaker conscious of the need to produce a subsequent sound which will require a relatively high degree of energy, such as an [s] beginning a stressed syllable, or a stressed vowel; this speculation rests on an inference from the claim defended by Ohala (*op. cit.*) of generally more or less constant pressure from the lungs, the inference being that if a constriction is maintained, pressure will build up behind it and be available to give greater force to a sound following the constriction (the mechanism behind oral stop articulations, of course).)

In spite of the relative insensitivity of nasal murmurs to the F2 of neighbouring vowels, in many consonantal contexts context-sensitivity is quite apparent. There are numerous cases of pre-velar [n] with a switch to an [ng]-like quality in the latter stage, usually with the switch itself registering in the acoustic record with a distinctive "fish-hook" transient, the constriction-position being altered while the velum remains lowered (figure 2.18, with the transient a little after two-thirds

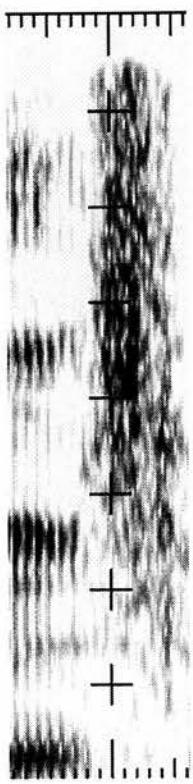


**Figure 2.15.** [n f] from utterance of "own photographs" (sc114)

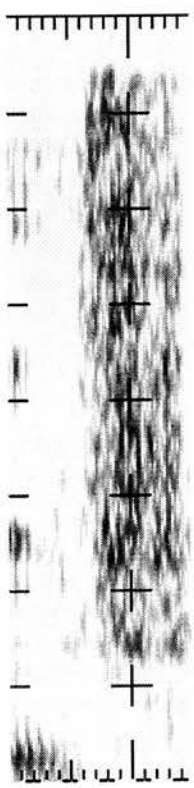


**Figure 2.16.** [n f] from utterance of "can finish" (sc127)

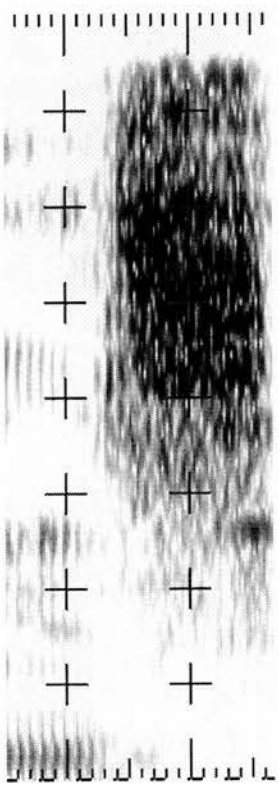


**Figure 2.17.** [n s] from utterance of "has been served" (sc146)

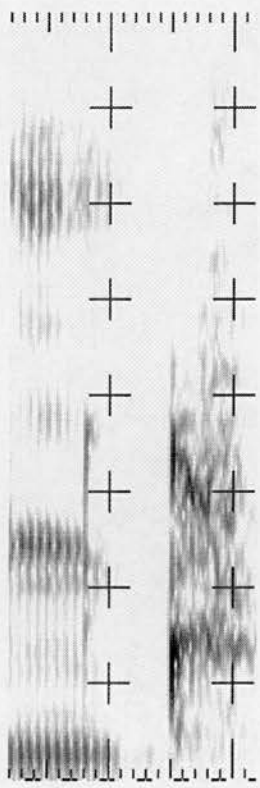


**Figure 2.18.** [n kc] from utterance of "an' cranny" (sc115)

of the way through the nasal, and 2.19, where the "fish-hook" is perhaps clearer, a little before half-way through the nasal). (If the reader attempts to produce utterances of this kind without allowing total assimilation of [n] to [ng], it may be possible to hear the tiny 'snap' which occurs when the front of the tongue detaches from the alveolar ridge, presumably the source of this transient in the spectrographic record.) Following [s], nasal murmurs typically begin only after an interval of silence, the formation of the occlusion for the nasal being accomplished before the resumption of voice, while before [-VOICE] stop-closures the latter phase of the nasal may be greatly weakened by early glottal opening for the stop.

As far as discrimination of nasals one from another is concerned, in GSW's speech the best cues are certainly the offsets of *preceding* and onsets of *following* sounds. Research into the relative importance of such cues, compared with information conveyed by nasal murmurs themselves, apparently yields contradictory

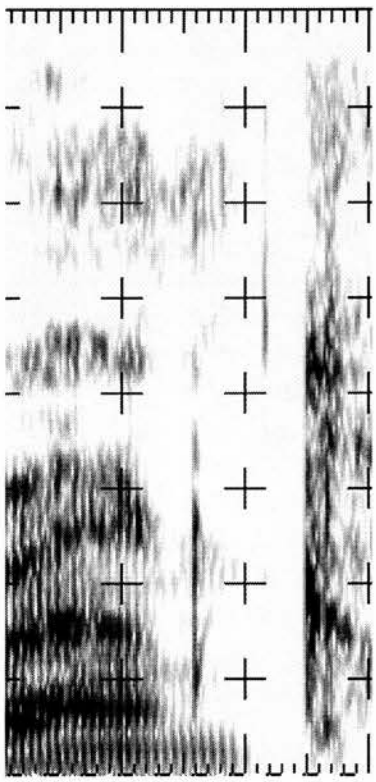


**Figure 2.19.** [aD1 n kc] from utterance of "ground coffee" (sc004)

results (Harrington 1987), transitions to and from the nasals being found more informative in some studies, and the quality of nasals themselves in others. In GSW's speech, though, it is almost everywhere the case that vowels to either side of [n]'s show just the same F2 movements as the same vowels to either side of other alveolars, and similarly for [m] and other labials, and [ng] and other velars, though of course the vowel spectra will often differ in the respective cases in respect of features such as nasality, breathiness and amplitude. Similarly, fricatives adjacent to nasals usually display characteristics associated with PLACE information similar to those displayed in the context of non-nasal consonants with identical PLACE features. Thus in GSW's case, the transitions to and from nasals offer significant cues to the identities of the nasals themselves.

Given a [-VOICE] stop and a following (usually but not necessarily homorganic) nasal, as in "at night" or "at most", it is necessary to lower the velum and close the glottis to effect the transition from stop to nasal (where the nasal is homorganic with the stop, this is all that is necessary). If the glottal closing gesture is relatively early vis-a-vis lowering of the velum, the stop-closure will be followed more or less immediately by some degree of nasal murmur (possibly rather weak initially), but if the glottal closing gesture is relatively late (vis-a-vis opening of the velic port), air may be driven out through the nose giving rise to weak high-frequency frication before any nasal murmur begins; in extreme cases, no nasal murmur may appear for the nasal at all, the articulators having already got into position for the sound that follows the nasal by the time voicing appears (figure 2.20). I explicitly labelled segments with late glottal closure as 'NPn' or 'NPm' (for (voiceless) Nasal Plosion), counting them as associated rather with the nasal than the preceding stop (thus treating the stop itself as 'unreleased'); the n/m subscript is made use of when the whole nasal is represented by such plosion, in order to provide contextual information for the onset of the following vowel, 'NPn' being used when the nasal is alveolar, and 'NPm' when it is labial. Voiceless nasal plosion is not in fact restricted to homorganic [stop nasal] pairs, and occurs in the data in [k n] (as illustrated in figure 2.20) and [t m] and [p m] pairs. ([ng] is of course not involved since it cannot follow a stop directly.) Segments or subphones marked 'NPn' or 'NPm' are not subject to any subsequent modification, and are treated as stand-alone subphones without any annotation

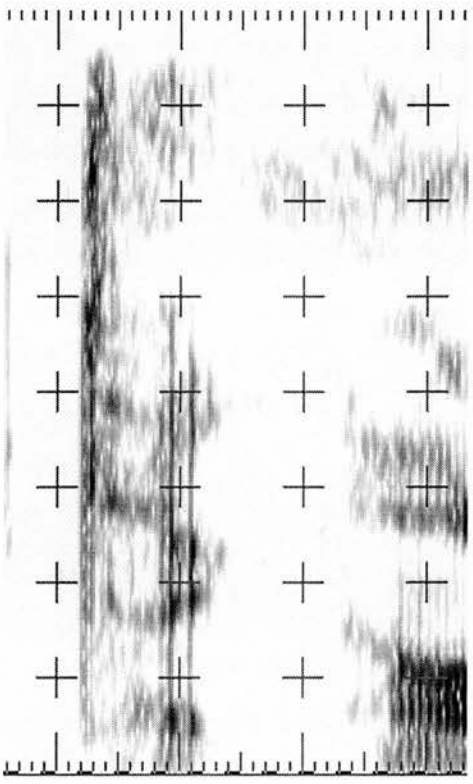


**Figure 2.20.** [t e GLkc NPn o] from utterance of "technology" (sc005), (where [GLkc] is a closure for a [k] that is preceded by glottalisation)

for right or left context. The n/m subscript is, however, later stripped away, nasal plosion tokens for all stops and all nasals being merged into a single class, [NP], because of extreme paucity of data, and because the acoustic effects of air rushing through the nose are not highly sensitive to the place of the oral constriction! Not much, of course, hangs upon the difference between, for example [ax tc NP n ai tc tb] and [ax tc n ai tc tb] and [ax tc NP ai tc tb] as different possible productions of the words "at night", but it is reasonable to suppose that introducing the explicit 'NP' label should help discrimination given the modelling assumption,[8] and we have here another example of a phenomenon connected with inter-articulator timing variation being handled easily within a linear framework.

[8]In the event, there was inadequate data for modelling [NP] in the transcription process.

Silent nasal segments or subsegments, other than those which are normally associated with a particular phonetic context such as following [s], are also labelled explicitly — as [no], [mo] and [ngo] — in order to preserve the sharpness and Normality of nasal subphone classes. These chiefly occur in the final one or two syllables of utterances, where overall signal amplitude is falling away, particularly before [-VOICE] fricatives or stop-closures, and before post-utterance silence.

There are syllabic [m]'s and [n]'s in the data, but they are few in number, particularly once the data is partitioned through the introduction of context-sensitivity. The syllabics are in *most* cases distinctive, but their distinctiveness is centred on their onsets, where we often find a more abrupt (if not necessarily earlier) appearance of robust nasal murmur than in the case of non-syllabics following the same consonants. So it was decided to model just the onset subphones of syllabics explicitly. Syllabics were labelled at whole-phone level with explicit syllabic labels ('nsyl', 'msyl', 'ngsyl'), but subsequent automatic procedures divided such phones into onsets retaining the syllabic label and offsets which were not distinguished from non-syllabic nasals. (Clearly, a subsequent automatic transcription involving the sequence [ nsyl (=onset) n (=offset] (where each subphone would be specific to a preceding or following phonetic context) would convert automatically to a syllabic [n] in the next stage of transcription.)

There may be varying degrees of nasalisation in sonorants preceding and following nasals, depending in large part on the relative timing of the velic gesture and the formation or release of the nasal stricture. More often than not, even where nasalisation is pronounced, characteristic formant patterns survive in the onsets or offsets of the affected sonorant. The data for nasalised laterals, [r], and [y] is very patchy, and this alone was a reason for not marking such cases explicitly in the labelling scheme. In the case of vowels, one reason for not explicitly marking nasalised tokens was just that although there were clear cases of nasalisation, nasalisation is a matter of degree and it was not clear how to implement a consistent labelling criterion. Another reason was that it appeared that the effect of nasalisation might be capable of being treatable as a variant on the basic formant pattern of the vowel in the given context, so that one could reasonably assume that the variability modelled in, e.g., the offsets of tokens of a given low back vowel preceding [n] would indeed be just (or largely) the variability that

| | |
|---|---|
| [n],[m],[ng] | piecewise left-/right-context-dependent |
| [nsyl] | ditto (right subphone losing syllabic marker) |
| [msyl] | ditto |
| [ngsyl] | ditto |
| [no],[mo],[ngo] | context-independent |
| [NP],[NPn],[NPm] | context-independent |

**Table 2.8.** Context-Dependency of Nasal Phones and Subphones

could be described in terms of degrees of nasalisation. There is one circumstance in which not explicitly marking nasalisation creates problems, however, and that is where the nasal is represented more or less solely by nasalisation, with no complete oral closure, though probably some gesture in the direction of such a closure. In the GSW-data there were only two or three cases of this kind in all, and in these circumstances a fudge was adopted, creating a false nasal segment in the most convincing bit of the acoustic record (clearly, the greater the gesture towards a closure, the more convincing the bit and the less the fudge), to save the chances of recovering the linguistic content and to provide an explanation at least for the nasalisation in the neighbouring vowel or vowels. But with more data of this kind, there would obviously be reason for working out a more sophisticated approach.

The context-dependencies of the various nasal phones and subphones are summarised in table 2.8.

### 2.5.5 Laterals

For accents such as those of GSW with a variety of lateral articulations, a simple left-right division and contextual annotation is insufficient to account for the acoustic variation found. One factor in this is devoicing, already alluded to in section 2.3 and to be discussed further below. Laterals of the various types found can be so dramatically different from each other that a simple left-right treatment that lumped all laterals for a given left or a given right context together would result in distributions that were poorly modelled by single gaussians; the first

part of a dark lateral following a high front vowel, for example, is quite different from the first part of a non-dark lateral following the same vowel (for GSW, perhaps the most salient difference is the gradualness of the change in formant pattern from vowel to dark lateral, and the abruptness of the change in formant pattern from vowel to non-dark lateral).

Non-dark laterals involve raising of the front part of the tongue toward the hard palate (together, of course, with tongue-tip articulation to form the lateral constriction), while in dark varieties the back of the tongue is raised in the direction of the soft palate. The parallel distinction between Russian palatalised and non-palatalised laterals is instructive, with the non-palatalised form representing an extreme form of the tendency evident in the English dark lateral. In discussing Russian laterals, Fant (Fant 1960) states that in the non-palatalised form the back of the tongue approaches the upper part of the pharynx, dividing the cavity behind the primary articulation into two, with a secondary point of articulation at or a little below the uvula, and that this results in a lowering of F2, with the lowering of F2 being linearly related to the narrowness of the pharyngeal pass. F1 is higher than in the palatalised laterals. In English dark laterals, the degree of backward displacement of the tongue-body appears to be less than in the Russian case, but in the English laterals too we find relatively high F1 and low F2 in the dark forms, while in non-dark forms F1 is normally lower, and F2 sometimes much higher than is ever seen in the dark forms, though in GSW's speech, the frequency of F2 is very variable and a great many tokens have F2 at quite a low level (around 1100 Hz). What is more distinctive in a large number of non-dark laterals in GSW's speech is the relative *weakness* of F2, regardless of its frequency. In the interests of trying to facilitate discrimination — and again given the simple statistical assumptions — I distinguished such laterals from those non-dark laterals that did not have a perceptibly weak F2. With the devoiced laterals (or devoiced parts thereof), this gives a four-fold division of laterals to be implemented in the manual labelling of the training-data:

- dl (voiced dark [l])
- cl (voiced non-dark [l] with relatively weak F2)
- l (other voiced non-dark [l])

- lo (devoiced [l], whether the lateral is dark or not, though only non-dark laterals were ever devoiced throughout)

The precise criterion for use of the 'cl' rather than the 'l' label was that F2 must be perceptibly weaker than both F1 and F3 (or than both F1 and combined F3 and F4 where the latter coalesced).

Apart from formant-frequencies and amplitudes, the basic distinction between dark and non-dark laterals is conveyed also by the manner in which the acoustic record evolves in the passage from a previous sound to the lateral and from the lateral to a following sound. In general, the transition is gradual for dark laterals when the neighbour is a vowel or non-nasal sonorant, but discontinuous for non-dark laterals in the same circumstances.

Turning now to the subphonic analysis, dark laterals are treated as consisting potentially of three subphonic moments, an onset, a sustained central part, and an offset, with the central part optional and not annotated for context (so that the central portions of all dark laterals from whatever phonetic contexts might contribute to its distribution). After [-VOICE] fricatives the onset of such a lateral may be devoiced, and in such cases the voiceless part of the lateral is manually labelled as such using the symbol 'dlo', the rest being labelled as 'dl'. Voiced non-dark laterals — both [cl] and [l] — resist simple generalisation with respect to their suitability for modelling with a simple onset-offset model: in many cases (figure 2.21, e.g.) there appears to be little influence from the F2 of adjacent vowels, with F2 more or less level throughout at a frequency not obviously connected in any way with those of F2 in the neighbouring vowels; in other cases, however, (figure 2.22 and 2.23) we can trace some measure of continuity between the F2 in the lateral and the F2 in the preceding and following vowels. On the other hand, even where F2 is level throughout, its frequency is highly variable across all tokens (ranging between about 1000 and 1700 Hz), the frequency in most cases not relating in any obvious way to the formant pattern of surrounding vowels. I am merely speculating in suggesting that this may be because of variation in the part of the tongue used to form the closure in individual cases, with consequent variation in the precise position and cross-sectional area of the constriction, and so too in the size and shape of the lateral outlet. A simple left-right division and contextual annotation was used for these laterals

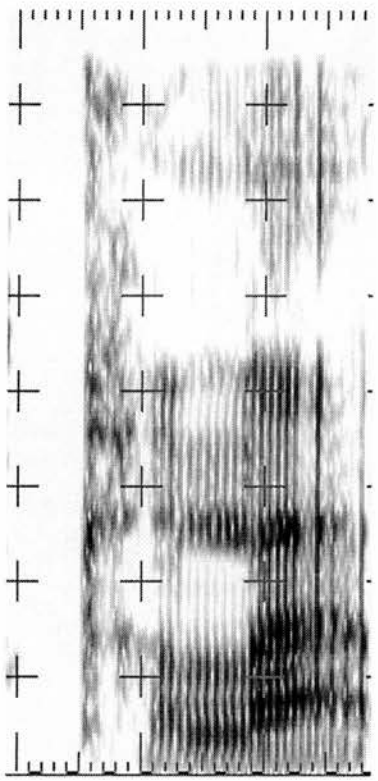


**Figure 2.21.** [k ax l a] from utterance of "collapsed" (sc150)

in the absence of any better ideas at the time, but the comments made above in connection with nasals, with regard to the possible appropriateness of mixture-modelling in abstraction from vocalic context, may apply equally well here.

Devoicing of laterals is more common than not after [-VOICE] fricatives (and to a lesser extent after [-VOICE] stops also), and dark laterals as well as non-dark laterals may be affected. After [-VOICE] fricatives, voiceless parts of laterals are explicitly labelled as such, using the 'lo' or 'dlo' symbol, and any remaining (voiced) part of the lateral is labelled with the appropriate one of the three available labels. To cater for cases of extreme delay in voice-onset, FURIDA allows non-dark laterals to be wholly represented by a single "onset" [lo] subphone annotated for left context; in these circumstances, a following vowel has *its* onset annotated as following on from a non-dark lateral (with no distinction of the fine species (allophonic type) of the lateral, or of its voicelessness). With larger

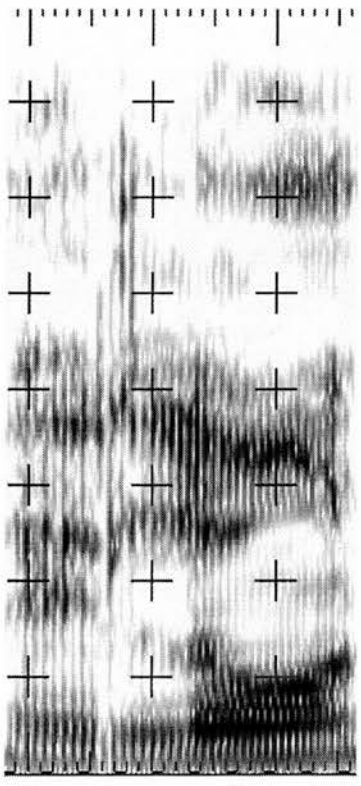


**Figure 2.22.** [ei cl oo] from utterance of "they launched" (sc011)

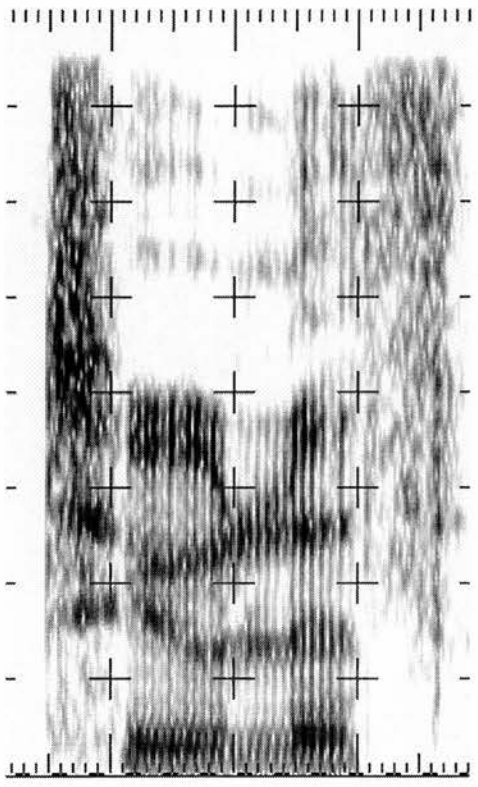


**Figure 2.23.** [t uuf l ax th] from utterance of "too lethargic" (sc168)

amounts of data, there would be an argument for dividing such wholly devoiced laterals into two subphones, since in the wholly devoiced case, there must be a relaxation of the lateral occlusion during the devoiced period, before the vowel articulation begins, while this need not be the case in partially devoiced laterals, where it is natural to suppose that the relaxation of the lateral stricture will be most evident during the final, voiced phase of the lateral. With limited data, however, it was necessary to accept an approximation here, with the consequence that the distribution for each devoiced lateral ([s_loB], [f_loB], etc.)[9] may include spectra closely or immediately preceding any of a variety of vowels; such spectra are likely to reflect acoustic effects of relaxing the lateral stricture and moving to the position required for the given vowel; in the single distribution for the left-context-sensitive lateral, however, these later spectra will constitute a minority (most laterals being voiced in their final stages), and thus not get the probability-score they deserve given the single gaussian used for modelling. With more data, it would be desirable to employ a right-context-sensitive right subphone, in addition to the left-context sensitive left subphone, for wholly devoiced laterals, in order to redress this deficiency. In the meantime, one can at least take some comfort from the fact that the features in common amongst spectra from different points in any devoiced lateral are probably more fundamental than the features in which they differ. Examples of devoicing are given in figures 2.24 to 2.26. Attention is drawn to the common occurrence of a thin transient marking the point where the lateral constriction begins.

After [-VOICE] stops, the place of articulation of the stop is obviously important for the precise realisation of the lateral. Alveolars almost always have a lateral release, while in labials formation of the lateral closure appears to be always more or less antecedent to or simultaneous with release of the closure. With velars, formation of the lateral occlusion is usually some tens of milliseconds later than release of the stop, typically resulting in a measure of shock to the system as the body of air released from the velar occlusion comes up against the new obstruction forward in the mouth, the lateral escape channels not being adequate for the free passage of the body of air released, so that transients may

[9] [s_loB] is the onset subphone of an [l] or [cl] following an [s], and may either be followed by a voiced offset subphone of [l] or [cl], or constitute the whole of the (fully devoiced) lateral.

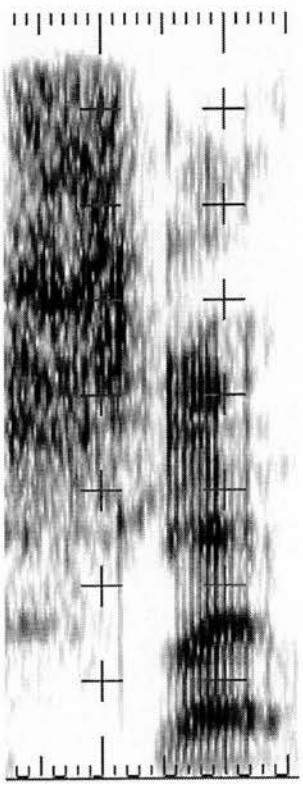


**Figure 2.24.** [s lo l e] from utterance of "slept" (sc168)

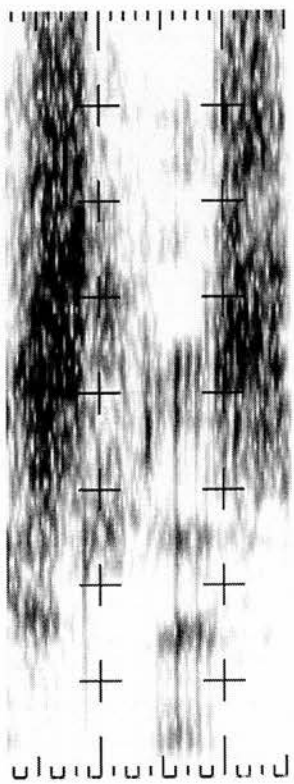


**Figure 2.25.** [s lo ax s] from utterance of "tasteless" (sc185)

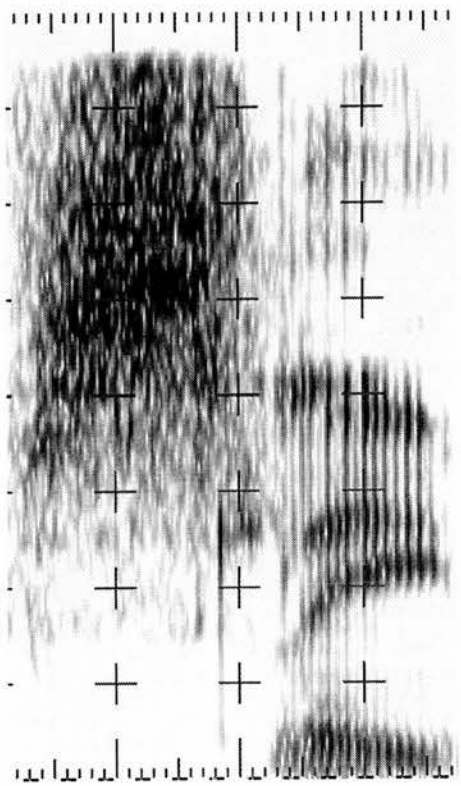


**Figure 2.26.** [s lo l ii] from utterance of "sleep" (sc187)

appear, probably marking the point where the released body of air meets the partial obstruction further forward in the mouth, with voicing appearing only some time after this point.

In the case of alveolars and labials, with very few exceptions it makes little sense to try to identify a point at which a stop-release ceases to dominate and a lateral occlusion takes over; in the alveolar case the stop-release typically *is* the lateral. In the case of velars, it is extraordinarily *difficult* in practical terms to find such a point when labelling. In all such cases a "coproduction" label is used to mark sections of the acoustic record where both the release of a stop and the formation of a lateral closure are in process, with the terminating boundary for such segments placed at the point where voice returns (a point which might occur during the lateral articulation or its relaxation, or not until the articulators are well into the process of shaping the following vowel). These phones — [tlb], [plb] and [klb] (i.e. simultaneous [t]- or [p]- or [k]-release and (devoiced subphone of) lateral) — are not subjected to any further expansion, and receive no contextual annotation, since their possible context of occurrence is highly restricted – a [tlb], for instance, being possible only after a [tc] and before either a lateral offset subphone or, in case of extreme delay in voice-onset, a vowel whose onset is annotated as following on from a non-dark lateral. Again a little recourse to approximations was necessary here, partly because of limited data, and partly because the voicing-change has a far more definite and easily recognised effect on the acoustic record than changes in tongue position during lateral frication: it is simply assumed that anything that happens prior to voice-onset can be modelled with one distribution, and anything that happens after with another. It must be said, however, that the use of single distributions to model these cases ([tlb], [plb] and [klb]) is again less than satisfactory, perhaps especially for [klb] with its high probability of transients or other indications of sudden turbulence in addition to the more uniform frication found throughout.

In a small minority of cases one does find stops released with a separate burst, prior to a lateral, and these are labelled in the conventional way (as [tb l] or [pb l] for example).

Finally (for this section) I turn to a discussion of syllabic laterals. The vast

majority of dark syllabics follow stops, with a few following fricatives. Acoustically, they do not appear *markedly* different from their non-syllabic counterparts, and were not given any specialised model, being labelled as [dl]'s along with other dark laterals. (Here as everywhere else, it was important not to partition the data unless there was a good acoustic reason for doing so). In fact the left contexts of the syllabics are peculiar to them (non-syllabics do not occur in the same contexts), so that no advantage accrues in terms of the amount of training-data available for particular models from merging syllabic and non-syllabic forms, as far as the onset subphones are concerned; but merging them assists in respect of the offset subphones, since both syllabic and non-syllabic forms may have identical right contexts, and acoustic differences are even less in evidence in the offsets of the two forms (for a given right context) than in the onsets. (The procedure used in connection with syllabic nasals — keeping the distinction between syllabic and non-syllabic forms as far as the onset subphone is concerned, but waiving it in respect of the offset subphone — could also be used here.)

Turning to non-dark syllabics, there are at most six in the data, and some of these are syllabic only in the sense that the phonological representations of the canonical forms of the words concerned have syllabic forms (as is the case for the lateral in the word "table"). With such a small amount of data, and such obvious room for doubt about what calling some of the tokens 'syllabic' actually meant, there was no point in trying to model syllabic non-dark laterals separately, and these laterals were treated in the same way as all other non-dark laterals.

The context-dependencies of the various lateral phones and subphones are summarised in table 2.9.

### 2.5.6 [r] and [w]

The label 'r' is used for the first sound of "red", and 'w' for the first sound of 'wed'. [r]'s are very regular in GSW's speech, with no flap-type tokens and all [r]'s showing some degree of F3 lowering. With intervocalic [r], the most obvious subphonic division is into the part of the [r] with falling F3 and the part with rising F3 (the subphonic boundary thus going at the lowest point in F3), these two subphones being quite realistically associated with the literal onset and offset

| | |
|---|---|
| [cl] after [lo] | right-context-dependent |
| [l] after [lo] | right-context-dependent |
| [cl] elsewhere | piecewise left-/right-context-dependent |
| [l] elsewhere | piecewise left-/right-context-dependent |
| [lo],[dlo] | left-context-dependent |
| [dl] after [dlo] | core context-independent, offset right-context dependent |
| [dl] elsewhere | onset left-context-dependent, core context-independent and offset right-context-dependent |
| [tlb],[plb],etc. | context-independent |

**Table 2.9.** Context-Dependency of Lateral Phones and Subphones

of the phone, if one assumes that F3 continues to lower as one proceeds with lip-rounding and retroflexion of the tongue, and begins to rise again as one begins to undo lip-rounding and retroflexion. Because [r] is so regular in GSW's speech, and the criterion for placement of the subphonic boundary so easy to apply, this division was executed manually in the labelling of the training data, using the symbols 'r1' for onset and 'r2' for offset, labels which subsequently get converted to encode left-context sensitivity (in the case of [r1]) and right-context sensitivity (in the case of [r2]), so that for example an [r] occurring between an [e] and an [ii], which would be manually labelled as [r1 r2], would be automatically converted thereafter to [e_rB r_iiA]. The precise placement of the subphonic border between an [r1] and an [r2] subphone is never subject to any further modification in these cases, since there seemed no need to improve upon it.

With post-consonantal [r], if voicing is present throughout, any section with falling F3 is labelled 'r1' as with intervocalics, and any section with rising F3 as 'r2' similarly. In the case of some consonants, the [r] may begin with voice, but with F3 already at a low-point and rising throughout, the tongue having already been put into a position of maximal retroflexion before the preceding consonant ceased to dominate the acoustic record; in these cases the whole [r] will consist only of an [r2] subphone. Such realisations are typical (though not inevitable) after [f] (the tongue being free to move during the [f] articulation, and oral pressure during the [f] not being so high as to cause a delay in voice-onset once the glottis closes), and after nasals; in the case of [ng], for example, it is

possible to effect retroflexion of the tongue-tip (and indeed lip-protrusion) while maintaining the velar constriction, so that at the release of the nasal only the offset of the [r] remains to be articulated (the [r2] may of course be nasalised in such a case, because of late velum-raising).

After [-VOICE] sounds involving high intra-oral pressure, such as [s] and stops, [r] may show varying degrees of devoicing. There do not appear to be tokens of [r] following [s] where the return of voicing is delayed beyond the point where maximal retroflexion is reached (this is probably also the point where a sudden drop in oral pressure occurs, making voicing easier if only the glottis is already in position for it. Hence [s] is never followed (in this data at least) immediately by [r2]; any section that precedes the beginning of [r2] is labelled as an [r1], regardless of whether it contains silence, a 'labial tail' to the [s]-frication, or normal voiced [r]. After [-VOICE] stops, presumably because of the high oral pressures generated, the delay in voice-onset may be so great that by the time voice appears the tongue and lips have gone through all the motions required for formation and relaxation of the [r] configuration, and are in process of preparing to form the configuration required for the following vowel. All intermediate degrees of voicing-lag also occur. To accommodate variation of this kind, the following strategy was adopted: where voice returned at or after the point of maximal retroflexion, (the low-point for F3), but before the vowel, all the subsequent section of the [r] was labelled as [r2] as usual, while any voiceless period extending from the end of the stop-closure up to that point was labelled with a 'co-production' label, 'trb', 'krb' or 'prb' (or 'Trb', 'Krb', 'Prb' for the unaspirated forms). These labels were used because it seemed to make little sense to try to distinguish a point at which the release of the stop ended and the [r] began, the articulators already being in process of fashioning the [r] at the moment the stop is released. In cases where voicing began prior to the point of greatest retroflexion, the voiced record therefore beginning with falling F3, the label 'r1' was used for just that voiced section ending with the low-point of F3. Hence a [t r ii] sequence (as for "tree") could be represented by any of the following, depending on the inter-articulator timing:

- voice-onset precedes maximal retroflexion – [trb r1 r2 ii] (cf. figure 2.27)

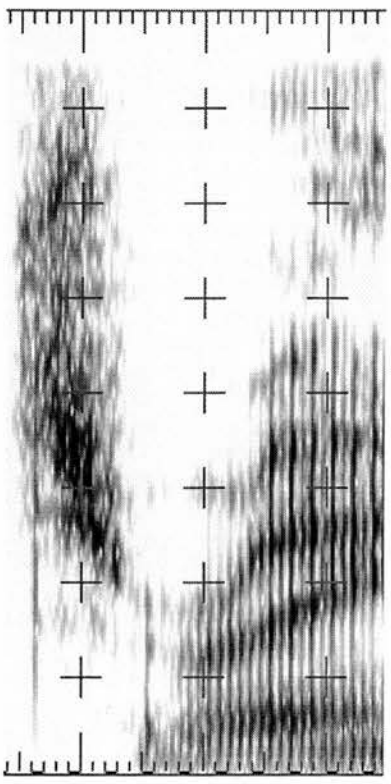


**Figure 2.27.** [trc trb r1 r2 ei] from utterance of "Mediterranean" (sc093)

- voice-onset simultaneous with or later than maximal retroflexion, but prior to vowel-onset – [trb r2 ii] (cf. figure 2.28)
- voice-onset delayed until vowel-onset – [trb ii] (cf. figure 2.29)

The 'xrb' segments ([trb], [prb], [krb], etc.) are stand-alone phones not given any contextual annotation and not subject to any subsequent analysis or modification. The lack of contextual annotation is, perhaps obviously, explained by the fact that the labels themselves largely define the context of occurrence, and the context of occurrence is obviously highly restricted: a [trb] can follow only a [t]-closure or at least a subphone involving transition toward a similar position to this, and can be followed only by either an [r1], an [r2], or a vowel with its onset marked as following on from an [r]-articulation. In the event of an [r1] following a

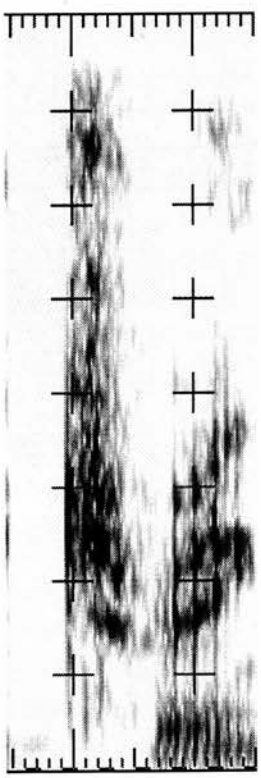


**Figure 2.28.** [trc trb r2 i] from utterance of "loitering" (sc008)

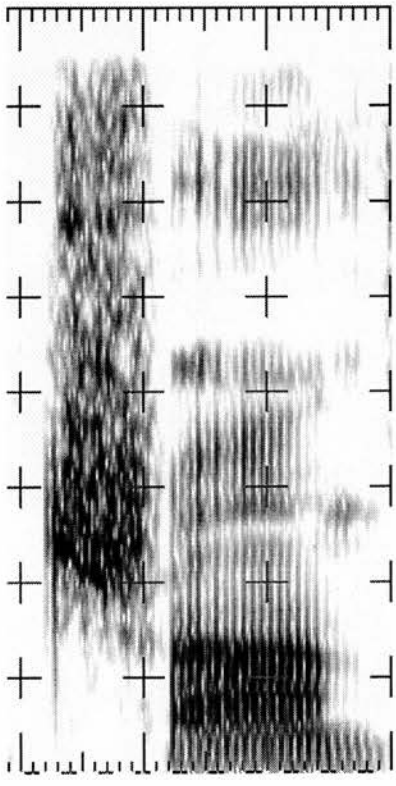


**Figure 2.29.** [trc trb aa n] from utterance of "entrancing" (sc160)

[trb] (and similarly for other POA's), the [r1] will subsequently have its left context annotated as [t], to facilitate merging with small numbers of tokens where voice-onset is very early and an initial brief stop-release is labelled separately as e.g. [tb], rather than as e.g. [trb].

Some subphonic analysis of 'xrb' segments would obviously be useful if some way could be found to do it, given that the distribution for the monolithic form ([trb], [prb], etc.) is liable to have minority representation for the offset spectra, cases of complete devoicing of the [r] being a minority of the total.

In the single case of alveolars, *closures* of stops with coproduced release and [r]-articulation are also labelled explicitly , as [trc] or [drc] (closures of unaspirates are also labelled as [trc]'s), for the following reason: particularly when fricatives precede such stop-closures, anticipatory lip-rounding so effects the offset of the fricative that the acoustic cues that normally reliably indicate the alveolar POA of the stop are replaced with a spectral trend in the fricative-offset easily confusible with a pre-labial or pre-velar; by explicitly labelling the closure as described, the fricative's right context is distinguished (as not simply alveolar, but alveolar and preceding an [r]), and the distribution for the 'normal' fricative-to-alveolar offset is not "corrupted", thus preserving good discrimination from the other cases. (This is the sole case in the system where context-sensitivity is actually accommodated to a distance of two phones.)

[w]'s are in most respects treated in a similar way to [r]'s, with 'w1' being used to label onset and 'w2' offset subphones, the subphonic division again being executed manually when labelling the training-data. Whereas in the case of [r] the low-point of F3 is taken as the dividing-point, in the case of [w] the division is put at a point where F2 is at a low-point, aligned with an energy minimum (typically no energy evident at all) in F3. [w]'s following some [-VOICE] sounds like [s] are again likely to have wholly or partly silent or voiceless onsets, and consonants which allow simultaneous lip-rounding and tongue-body movement or which have compatible tongue-body specifications, ([dl] (dark [l]), [m], [n] and [ng] are all examples) are capable of being followed directly by offset-only [w]'s, i.e. by a [w2] subphone.

The only difference between the treatment of [w] and the treatment of [r] arises in connection with [w]'s that follow released stops; in most cases involving

[r], a stop-release was not separated from a voiceless [r], but — largely for no more than historical reasons of no theoretical interest — this pattern was not followed in respect of [w], so that there are no 'twb' labels to parallel the use of 'trb' labels, for example. Generally it was straightforward to label stop-releases separately from following [w]'s, either because a silent period intervened (in such a case, the silence was counted as part of the [w1]), or because voice-onset marked a clear point for putting the beginning of the [w1] or [w2]. It would probably be preferable, though, if only for simplicity's sake, to have a uniform policy for the two cases, since [r] and [w] are so similar in every other respect.

### 2.5.7 [h]

The glottal fricative or approximant [h] (the first sound in "he") is often said to have no specification of its own for any supraglottal feature (though free passage for airflow through the mouth is clearly required). Certainly when intervocalic, [h] normally shows a formant pattern that 'interpolates' between the formant pattern of the preceding and the formant pattern of the following vowel as the tongue is repositioned in more or less complete indifference to the ongoing glottal articulation. Intervocalic [h]'s, then, should ideally be modelled triphonically, rather than in piecemeal left-right fashion, since even in the earliest part of a [h], the formant pattern of the following vowel will have some bearing on the formant pattern evident in the [h], though obviously that influence will become more telling as the [h]-articulation continues. With insufficient data for good triphone modelling, I attempted to get as close as was feasible to it as follows: intervocalic (and indeed all post [+VOICE] ) [h]'s were labelled at phone-level as [PVh] ('PV' for "post-[+VOICE]") and subsequently divided automatically into two subphones, each subphone annotated in such a way as to indicate the overall trend in F2 in the passage from the vocal tract configuration (VTC) for the preceding to that for the following vowel. Thus, if a high front vowel preceded the [h], and a low back vowel followed, the annotation for *both* subphones would indicate a passage from a high frequency F2 to a low frequency F2, with, obviously, the first subphone modelling the phase of the [h] during which the first half or so of the articulatory change took place, and the second that in which the

rest of the change took place. Because of limited amounts of data, F2 levels were counted as belonging to one of three levels, high, mid, and low, so that slopes were either level, rising, or falling, giving nine basic patterns, each pattern being modelled in two phases, early and late. As an example, a [h] occurring between an [oo] (the wowel in "saw") and an [ii] (the vowel in "see") would be assigned to the category of tokens with a 'low to high' slope in any second formant that might be discernible. In the initial step of the process of automatic expansion of the simple manual label 'PVh', the phone is divided 50:50 into a left subphone and a right subphone, but both preceding and follwing vowels are taken into account in assigning names to these two subphones; the left subphone in this example is named as a token of [h0L2HX] and the right subphone as a token of [h0L2HY], where the 'L2H' indicates merely the low to high slope in F2, and the 'X' is used to indicate the left hand, and the 'Y' the right hand subphone. Further automatic processing leaves room for adjustment of the precise placement of the subphone boundary, but that is to anticipate the subject-matter of the fourth chapter. Variation in the degree of voicing in intervocalic [h]'s is left in FURIDA as a form of 'residual variability', but in fact is poorly handled in this way given the statistical modelling assumption. Some [h]'s are as strongly voiced as vowels, while others are wholly voiceless, and this therefore appears to be yet another case where a bi- or multi-modal gaussian would be more effective for modelling-purposes.

In the case of [h]'s that follow [-VOICE] sounds, it seemed to be a more or less general rule, for GSW's speech at least, that articulatory repositioning for the following vowel or [y] actually began in the offset of the consonant *preceding* the [h], so that the [h] itself was characterised more or less throughout by a level formant pattern equal to that of the following vowel or [y]. These [h]'s were treated as follows. They were labelled at phone-level as [PNVh], and subjected to no further division, but converted automatically to a right-context-sensitive form with annotation for the broad frequency-level of the F2 of the following vowel or [y]; it is possible, thus, to say that these [h]'s were modelled with a single subphone representing something at least very close to a steady state, any glottal gesture and supraglottal repositioning having already been more or less accomplished at the beginning of the segment, and no supraglottal change being

required in the transition to the following sound; some indeterminacy surrounds the question of an "offset" in respect of the glottal state, but low oral pressure means that the passage from open glottis to the setting for voice is quite rapid, though in any case some breathy voice is possible during these [h]'s, though less so than in the case of [PVh]'s.

There is one case of inter-articulator timing-variation affecting [h] which is worth describing here. When [h] precedes an [ou] diphthong (the diphthong in "go"), in GSW's speech at least, it can happen that the timing of voice-onset is so delayed that the oral tract has already been put through the configuration required for the first element of the diphthong (labelled as [axD1] in this work) and is at some point along the way to the achievement of the configuration for the second element (labelled as [axrD2] or [uumD2] in this work) by the time [h] ceases to dominate the spectrum.[10] Given this state of affairs, the labeller has to indicate whether a D1 element is present at all; if voice returns with some part of the D1 element (however short) remaining to be produced, the labelling shows [h axD1 axrD2], but if the onset of the D2 element is already in evidence at the voice-onset, the labelling shows [h axrD2]. The subsequent conversion to subphones follows accordingly, and the sequencing-rules have to allow the [h] offset to be followed by any of the onset, core, or offset of the D1 element if one is present in the manual labelling, and directly by the onset of the D2 element if no D1 element is present. It is judged extremely improbable that one would find voice-onset delayed so greatly that a core or offset D2 followed the [h] offset directly.

[10]As will be explained further below, the diphthong [ou] is modelled as if it were two vowels in sequence; the similarity of the first part of [ou] to schwa ([ax]) in the speech of GSW prompted its labelling with the symbol 'axD1', where 'D1' is mnemonic for 'first element of a diphthong'; the second element of this diphthong is sometimes realised with a quite [uu]-like quality (as in the vowel of "two"), in which case it is labelled as [uumD2], and sometimes realised by merely adding a measure of lip-rounding to the articulation used for the first element, in which case the symbol 'axrD2' is used, the 'D2' in both these cases being mnemonic for 'second element of a diphthong'.

### 2.5.8 [y]

[y] (the first sound in "yes") functions rather like a vowel in most respects; it may have a sustained (if usually rather brief) 'core' state, and is therefore modelled as consisting potentially of three subphonic moments (onset, core and offset), with the onset conditioned by left and the offset by right context, and the core subphone treated as context-neutral (though this as an approximation more or less forced by data limitations), so that core subphones from [y]'s in all contexts are modelled with a single distribution. Only phone-level labelling with a single label 'y' is involved in the preparation of training-data.

As with other [+VOICE] sounds, when a [-VOICE] sound precedes, the onset of voicing may lag behind supraglottal articulatory changes; the formation of the palatal stricture may already be accomplished to varying degrees, and may even be in process of being relaxed, by the time that voice appears. I would have liked to handle such variation in a way similar to that used for laterals and [r], but two problems prevented this: one is that it is not always clear how much delay there is (the palatal stricture is sometimes easily discernible in the frication or aspiration associated with the previous sound, but not always so, or a silence intervenes and any boundary-placement would be arbitrary), and the other is that of likely data-shortage problems for any [yo] (or [tyb] etc.) subphones.[11] As far as I can recall there was no case of an entire [y]-articulation (from beginning of onset to end of offset) taking place during the voice-lag, so the lack of a special symbol or symbols was at least not critical. The variation was handled by allowing variable subphonic realisations of [y] following [-VOICE] sounds, assigning all devoiced regions to the previous phone even where [y]-like formants were visible, and making the [y]-segment begin with the beginning of voice. Thus after the offset of a [-VOICE] phone identified as having [y] as its right context, we could have either an onset of a [y] (with appropriate left context), or a core [y] subphone, or an offset of a [y] with any right context.

[11]The 'yo' symbol, had it been used, would have indicated a devoiced [y] phone or subphone, and the symbol 'tyb' a stretch of the acoustic record dominated by release of [t] with the articulators already well in position for [y].

| Vowel | Example |
|---|---|
| [ii] | "ease" |
| [rii] | "reason" (see text) |
| [i] | "is" |
| [ir] | "is" (see text) |
| [e] | "fez" |
| [oe] | "well" (see text) |
| [a] | "Baz" |
| [ax] | "the" |
| [lax] | "bird" |
| [aa] | "par" |
| [uh] | "much" |
| [o] | "not" |
| [u] | "put" |
| [oo] | "caught" |
| [uuf] | "newt" (see text) |
| [uum] | "rune" (see text) |
| [uu] | "school" (see text) |

**Table 2.10.** Labels for Monophthongal Vowels

### 2.5.9 Monophthongal Vowels

The phone-level labels for monophthongal vowels are listed in table 2.10. Monophthongal vowels are labelled at phone-level, and subsequently expanded into a minimum of two and maximum of three subphonic elements as with [y]. All core or steady states for a given vowel class are modelled using a single distribution; in the case of vowels, this is not generally an unappealing practice, given that the identity of the vowel is crucially associated with the formant pattern of any steady state it has. In the majority of cases basic contextual variability of vowels is accommodated by representing them in terms of left-context-sensitive onset, context-neutral core state, and right-context-sensitive offset. In a small number of cases, however, this is quite inadequate.

One case requiring special treatment is /uu/ (the vowel in "newt", "rune", and "school"). Perhaps because of the irrelevance of the FRONT-BACK distinction to English /uu/, or perhaps because GSW had one Scottish parent and one English

parent, (and /uu/ is normally very fronted in Scottish English), F2 appears over a great range of positions in GSW-ese: in most tokens it is as high as in [i] or [ii] (around 1700 Hz) (in these cases, F3 is noticeably lower than in [i] and [ii]), in a fair number of tokens it is more intermediate (around 1400 Hz), and in a small number of cases it is quite low (around 1100). While the phonetic context seems to have a bearing on the F2 level found in any particular case, the influence is not a simple or regular one; in any case, the data for particular contexts is very patchy, and it was decided to "quantise" into three cases of fronted (F2 above about 1600 Hz), intermediate (F2 between 1300 and 1600) and back (F2 below 1300), using 'uuf', 'uum' and 'uu' as the respective "allophonic" symbols. Phones of the three types are subsequently subject to subphonic analysis and contextual annotation in the same way as all other vowels.

The /e/ vowel (the vowel in "fez" and "well") has a realisation with markedly lowered F2 when occurring after [w] or before [dl] in GSW's speech, so that a single model for steady-state subphones from [e]'s in this and [e]'s in other contexts would have had to be bimodal. I therefore labelled the [e]'s in these contexts separately as [oe]'s ("open [e]'s"), and otherwise treated them in the same way as all other vowels.

[r]'s tend to have a very persistent effect on following [ii]'s (and [iiD1]'s – see the next section on diphthongs), so that even in vowels with sustained core states F3 remains significantly lowered. For this reason, allophones were introduced in this case also; any [ii] following [r] was labelled as [rii], and any [iiD1] as [riiD1].

The coding[12] for FURIDA took about two years, and began from a level of phonetic understanding which remained largely static until the programming was nearly complete. Consequently, FURIDA embodies some phonetic ideas which I can now see are mistaken. One of these is reflected in the handling of schwa. Work by Bates (Bates 1995), Kondo (Kondo 1995) and others has shown that the acoustic realisation of schwa is dependent in all cases on the phonetic context in which it occurs; the frequent occurrence of schwas with level F2 at around 1500 Hz for a typical male speaker is not the reflection of a 'normal' or 'canonical' target-value for schwa, but simply a result of the commonness of schwas in contexts with

[12]By coding is meant here writing of C programs.

alveolars both before and after. In ignorance of the true situation, I took schwas with level F2 at around 1500 Hz as representing a kind of norm, but could not help noticing the large number of "fronted" schwas with high F2, as well as the large number of cases of [i] which seemed to be reduced in the direction of schwa. I tried to reduce the variability as between the fronted and non-fronted cases by introducing an 'ir' label and labelling "fronted" schwas and reduced [i]'s as [ir]'s and all other schwas as [ax]'s, without any clear understanding of how it was that certain schwas appeared like back vowels, or any strategy for dealing with them. Schwas, like intervocalic [h], ought to be modelled using triphones.

Returning to the main theme of monophthongal vowels in general, onsets are made optional after [-VOICE] phones (the reasons for this should be clear from the discussion of [-VOICE] stops and fricatives in preceding sections: during stops, the repositioning of the tongue that is required for the following vowel may be accomplished wholly during the release-phase of the stop; in the case of fricatives, it is often possible to maintain the stricture while beginning tongue-body adjustments for the following vowel, while the brief but real amount of time required to change from voiceless to voiced articulation also helps to make it possible for a vowel to begin with its 'core' quality following a [-VOICE] fricative). It is therefore possible to progress directly from a [-VOICE] phone with its offset marked as preceding vowel X to a core subphone for X. Core states are also optional for all vowels, though that option is not very commonly taken up by vowels other than schwa, [i] and [u]. Monophthongal vowels with no core are constrained to have an onset even after [-VOICE] sounds (the occurrence of a vowel which consisted simply of an offset may not be an incoherent one, but there were no obvious signs of such things in the data for GSW, and I am doubtful about whether I have seen such cases anywhere). All vowels are required to have an offset in FURIDA, though there are reasons for thinking that this stipulation should be relaxed in some cases. In the cases concerned the offsets can sometimes be so brief as to be unreliable as a source of contextual information. This is so with the vowels [ii] and [uuf] before alveolar and velar stop-closures, and also sometimes with low back vowels before non-dark laterals. The problem is exacerbated by the use of border-straddling classes (see Chapter 3, section 5) to model the boundary-regions between phones.

| Diphthong | Modelled as |
|---|---|
| [au] (as in "now") | [aD1] and [axD2] or [uumD2] |
| [ou] (as in "go") | [axD1] and [axrD2] or [uumD2] |
| [i@] (as in "here") | [iiD1] and [axD2] or [irD2] |
| [e@] (as in "there") | [eD1] and [axD2] or [irD2] |
| [E@] (as in "care") (close [e]) | [eHD1] and [axD2] or [irD2] |
| [u@] (as in "poor") | [uD1] and [axD2] |

**Table 2.11.** Two-Phase Diphthongs

Offsets of vowels before [-VOICE] fricatives provide another instance where variation in inter-articulator timing can have significant acoustic effects. The relative timing of tongue-readjustment and of relaxation of the glottal state for voicing determine the degree to which increasing breathiness appears in the offset of the vowel, and the degree to which voice persists into the onset of the fricative. In extreme cases there may be complete loss of voice before the consonantal stricture is effected, sometimes even before the tongue has begun to abandon the steady position for the vowel; these cases typically occur when the vowel is the last vowel before a major clause boundary (and the fricative is part of the same word as the vowel). These vowels are ill-suited to the simple onset-core-offset modelling used in this work, but [at the time of writing] I have not found time to implement a more appropriate strategy for dealing with them. Either explicit annotation of the vowels with respect to clause-position, or modelling of the vowels with a larger number of subphones, or mixture-modelling, or some combination from these, would appear to represent promising avenues to improvement.

### 2.5.10 Diphthongs

Non-rising diphthongs are modelled as sequences of two vowels, as shown in table 2.11. In labelling these diphthongs, the boundary is placed half way along the transition between the two component vowels. In the event of one of these diphthongs being realised in a 'curtailed' form, with only the first 'vowel' being produced, the second being skipped, only the first 'vowel' is labelled in. Thus

| | |
|---|---|
| [ai] | [aaD1 aa2ii iiD3]<br>or [aaD1 aa2i iD3]<br>or [aaD1 aa2e eD3]<br>or [aaD1 aa2ax axD3] |
| [ei] | [eD1 e2ii iiD3]<br>or [eD1 e2i iD3]<br>or [eD1 e2ax axD3] |
| [oi] | [ooD1 oo2ii iiD3]<br>or [ooD1 oo2i iD3]<br>or [ooD1 oo2ax axD3] |

**Table 2.12.** Manual Labels for Maximal Realisations of Rising Diphthongs

a rapid utterance of the words "now that" in which the diphthong of "now" was realised without execution of the articlation required for the second 'vowel', would see the "now" labelled manually as [n aD1]. Each of the individual vowels is treated in the same way as any monophthongal vowel in subsequent processing. With the exception of the possible omission of [axD1] following [h], explained above, every 'D2' vowel must be preceded by a 'D1' vowel.

The remaining three (rising) diphthongs represent rather more difficulty, and are labelled at subphonic level as follows: the period preceding any glide from a first element to a concluding state is labelled as a 'D1' element, and any glide is separately labelled as such. The boundary between D1 and glide is placed at the point where F2 begins to rise. Any final state that is sustained even momentarily is labelled as a 'D3' state. As is well known, the final state in these diphthongs has quite variable realisation, and the manual labelling is made sensitive to this. Table 2.12 shows the possible manual transcriptions of these diphthongs when they are fully realised (with initial, glide and final state):
In individual cases various degrees of curtailment are possible to the maximal realisations of the diphthongs given in table 2.12: frequently in the case of the /ei/ diphthong, and occasionally in the case of the /ai/ diphthong, the D1 element is omitted, the diphthong beginning with a glide. In the case of all three diphthongs, the D3 element may fail to be articulated at all; where this happens, but the glide does get articulated, the left context of any sound following the diphthong

| manual labelling | converted to |
|---|---|
| [w aa2ii] | [...w_AIgl1 CONS_AIgl2] |
| [f aa2i] | [...f_AIgl1 CONS_AIgl2] |
| [m e2ii] | [...m_EIgl1 CONS_EIgl2] |

**Table 2.13.** Automatic conversion of post-consonantal glides

is equated with the sound referred to in the second part of the glide label ([ii] for [a2ii], [i] for [aa2i], and so on). Where the glide too fails to be articulated, so that the diphthong is represented by its D1 element alone, the whole diphthong is equated with the corresponding monophthongal vowel ([aa] for [aaD1], [oo] for [ooD1], [e] for [eD1], with an offset leading to the following sound. It follows from the labelling scheme that D1 elements that get labelled as such never have offsets (being followed immediately by glides), while D3 elements never have onsets (beginning as they do at the precise point where F2 ceases to rise and is sustained at a particular frequency).

Subsequent to manual labelling, further subphonic resolution occurs automatically to distinguish an onset from a core state in the D1 element, and to distinguish a core from an offset in the D3 element. In all circumstances, D1 elements that follow [-VOICE] sounds have optional onsets, and where an onset is present have optional core. Glides are not subjected to further subphonic analysis when they are preceded by a D1 element, but when preceded by a consonant they are divided into two subphones, as the examples in table 2.13 illustrate.

The reason for the subphonic expansion of the post-consonantal cases was a desire to reduce the chances of insertions of spurious sliver-length glides in automatic transcription, given the unusually weak sequence-constraints in this area (with diphthongs capable of being realised by a glide and nothing else). Insistence on a two-subphone realisation of the post-consonantal diphthongal glides was an early and rather weak measure in this direction (it would be slightly more difficult to find two consecutive slivers of the acoustic record that could (mistakenly) be classified as belonging to a diphthongal glide, than it would be to find a single such sliver). Originally even the glides that were preceded by D1 elements were subjected to subphonic resolution in the same way, as it was

| manual labelling | converted to |
|---|---|
| [aa2ii] | [AI_IIgl] |
| [aa2i] | [AI_Igl] |
| [aa2e] | [AI_Egl] |
| [aa2ax] | [AI_AXgl] |
| [e2ii] | [EI_IIgl] |
| [e2i] | [EI_Igl] |
| [oo2ii] | [OI_IIgl] |
| [oo2i] | [OI_Igl] |
| [oo2ax] | [OI_AXgl] |

**Table 2.14.** Conversion of post-D1 glide labels

originally thought that the sharper distributions that result from such an analysis would be wholly desirable. In the event, it turned out that this very sharpness might be responsible for a problem that was never really solved — the tendency to misrecognise vowels with rising F2 in their offsets, especially the vowels [e], [aa] and [oo], as diphthongs. Returning to a single-subphone realisation for the diphthongal glides that followed D1 elements was motivated by a suspicion that their sharper distributions might be giving them unfair advantage over vowel-offsets (which are each modelled as a single subphone) in the scoring that is at the basis of the classification system used for automatic transcription (this will be explained in full in due course). The reversion to single subphones would have been taken up for the post-consonantal glides too had it not been for the problem described above. Some merely trivial translations of the labels for post-D1 glides take place as shown in table 2.14. Many of the details of the conversions, as indeed of those for post-consonantal glides, are connected with the need to get round data-shortages.

One problem with conferring right-context sensitivity on those glides that are not followed by a D3 or some other vocalic element is that consonantal context does not seem wholly to *determine* the specific target reached in the glide, though it certainly exercises considerable influence. The data for GSW at least suggest a conditioning rather than a determining effect: an [ax]-type target is most likely with labials [v], [f], [b], [w], [r] and with [l] and [dh], but also possible with [h],

[s] and [n]; an [i]-type target is more likely with alveolars [s], [z], [t] and with [sh] and [l], but also possible with [g], [v] and [h]; and an [ii]-type target is usual with velars, but also possible with [t], [n] and [jh]. Patchiness of data makes it necessary to work with generalised contexts, and the fact that we have only tendencies means that in some cases we will find acoustic characteristics suggestive of an inappropriate consonant-group. Certainly for the post-consonantal glides the conversions mean that we sacrifice the specific information about acoustic 'targets' encoded in the manual labelling.

## 2.6 Elision, Assimilation and Fusion

Before leaving the subject of labelling, it is necessary to explain the policy adopted for dealing with cases of elision, assimilation and what is here called fusion.

First, a brief word on what is usually called elision. Browman and Goldstein (Browman & Goldstein 1990) may or may not be right about there always being residual gestures for segments often described as "deleted" (it seems unlikely to me to be a yes-or-no question), but from the labeller's point of view, what matters is whether there is any acoustic or auditory evidence, and moreover evidence that is so great as to seem worth taking note of. If the words "fast buck" are pronounced as far as one can tell as [f a s b uh k], with the [s] clearly showing a labial rather than an alveolar trend in its offset, then the utterance must be segmented and labelled accordingly. (This is a fairly trivial example of the flexibility of the relationship envisaged between acoustic segments and the conventional linguistic representation sought as the final goal.)

Assimilation (in the present context this will always mean PLACE assimilation) may mean nothing more than an adjustment in the usual place of articulation (POA) of the sound concerned, in the direction of the POA of (usually) the following sound, as for example with dental articulations of [n] before [th]. In some cases, the adjustment amounts to a wholesale adoption of the POA of the following sound, and in certain circumstances, as with [n] fully assimilated in PLACE to a following labial or velar, the result is a sound that is acoustically indistinguishable from one of the sounds that contrasts with the sound assimilated (an [m] or [ng] in the example given). The policy adopted in labelling with respect

to assimilations of the former kind (dental [n] etc.) is to ignore it, on the basis that the characteristics of the *onset* will tend not to be dramatically different from those of an unassimilated sound in the same left context, and can be handled as residual, while the offset characteristics will be captured by right-context sensitivity in the usual way (again the variation in degree of any assimilation being handled via the basic statistical modelling). With respect to assimilations of the second kind, however, where the assimilation is total (as revealed in the formant-transition at the end of the preceding sound), labelling is in accordance with the output of the assimilation, rather than the input, so that "in between" with full assimilation of the [n] will be labelled as [i m b ax t w1 w2 ii n]. But where, in assimilations of this same second kind, the change in PLACE occurs only some time after the beginning of the sound concerned (as was described in section 2.5.4 in the particular case of nasals), the labelling is in accordance with the input. We perhaps have here yet another instance where mixture-modelling would seem desirable; for example, some pre-velar [n] offsets will have [ng]-like murmurs because of late assimilation, while others will retain [n]-type murmurs to the end.

I did not treat cases of voicing and devoicing as cases of VOICE assimilation because of the wide variety of factors that may be responsible for these phenomena, and because the VOICE feature in [+VOICE] sounds in English is not carried simply by the presence of physical voice. I therefore labelled in accordance with the phonological feature rather than the acoustic characteristics as long as there was *some* cue in the acoustic record that the speaker was at this point distinguishing the sound as [+VOICE] or as [-VOICE], e.g. by lengthening a preceding vowel. It may, of course, occur that a [+VOICE] fricative preceding a [-VOICE] sound may be devoiced in all or part of its frication (and of course the convention is to mark the segmental boundary as beginning where the frication begins), and where the following [-VOICE] sound is itself a fricative, and of identical PLACE, the acoustic record for the two fricatives will show a seamless continuum, and a convention is required for such cases to determine consistent placement of the boundary. In this work, where the [-VOICE] fricative was either unstressed or only lightly stressed, the frication was apportioned 1:2 between the [+VOICE] and [-VOICE] pair, but, depending on the degree of stress of the

[-VOICE] fricative (or of the syllable it began), this proportion could be extended up to 1:4. We may also find cases where the second fricative differs in PLACE and the first fricative assimilates in PLACE to it. In these circumstances, the same procedures as outlined above are used; if, for example, the acoustic record for an utterance of "was sure" indicates that the articulators passed directly to a palato-alveolar POA on leaving the schwa, giving rise to frication which has from the outset a palatal rather than alveolar quality, then the fricative is labelled as a [zh] (the VOICE feature being decided as indicated previously), otherwise as a [z].

One step further than assimilation takes us to *fusion*, where both of a pair of sounds get produced as a single hybrid sound with some of the features of both, as for example when the utterance "How's your head?" gets produced as [h au zh ax h e d]. Trying to impose 'z' and 'y' labels on the spectrographic record in these circumstances is unappealing, and it seems more sensible to cast some additional burden on later phases of the recognition process and label in such a way as to exploit the acoustic similarity to the sound whose symbol we are borrowing. There is a point of doubt, though: in the example given, we use the symbol for one of the contrastive sounds of the language ('zh') to label the hybrid, and it is certainly the case that this kind of labelling of the commonest forms of fusion of this kind — forms like "d'you" produced as [jh uu] or [jh ax] — is highly acceptable (institutionalised, so to speak) (Barry & Fourcin 1990). But is it the case that all fusions of this kind can be represented by "borrowing" a symbol for one of the contrastive sounds of the language? There is obvious room for doubt.

There are not a great many examples of fusion in the data; two of them are fusions of [ii] and [w] to produce a sound I have labelled as [uuf], as in "We were" produced as [w uuf ax] and "She was" produced as [sh uuf ax z]; again, imposing a [ii w] label-sequence on the data in these circumstances seems quite unappealing. In one or two cases some disturbance of the normal articulatory sequence — maybe from excessive haste in delivering the scripted sentences — resulted in articulations of [y] in which only the onset was produced or at least audible, with the result that [ii] rather than [y] seemed a better transcription of the resulting sound: hence "... as you should know by now" produced as [ax z ii

sh ax d ...] and "invaluable" as [i n v a l ii b dl].

Clearly there are styles of speech where processes of this kind become so pervasive that the chances of recognising the linguistic content from phonetic transcriptions of the kind advocated here become rather small. Well-known phonological rules can be used to cope with common forms of elision, assimilation and fusion (Oshika *et al.* 1975), but getting to a stage where we had rules to accommodate all eventualities would be no small undertaking. Consider, e.g., the rules that would be needed to get word-sequences from the following transcriptions of real (spontaneous) speech:

- "What do you have?" as [w ax tflap ii h a v] (Canadian Map Task Corpus)
- "following it" as [f o l i ng i t] (ditto)
- "Which side?" as [w i t s ai d] (ditto)
- "Manchester United" as [m ã sh T ax y uuf n ai GL ax d][13] (Terry Venables, ex England football coach)

Some comfort may be taken from the fact that it seems unlikey that people would address a computer in such styles of speech (at least until the day arrived when people found you could do so without problems!).

## 2.7 Glottalisation – the Segmental Framework tested to the limit

Glottalisation presents perhaps the most difficult problem of all for the segmental treatment of speech, but I shall endeavour to show that the problems can all be overcome. As it is not possible to discuss the issues involved without considering the linguistic functions of glottalisation, that is where this section will begin.

Glottalisation has two primary functions in GSW's speech. It may be used as a substitute for a [-VOICE] stop or affricate-closure, and it may be used to highlight juncture, between either a consonant and a vowel or between two vowels,

[13]The tilde over the [a] indicates nasalisation.

with a view to attracting attention to and so emphasising the lexical element beginning with the (second) vowel. For convenience I shall refer to these two broad categories as STOP SUBSTITUTION and JUNCTURE EMPHASIS. (The fact that the speech-data used in this work is scripted means that it is not necessary to consider a 'use' of glottalisation that would certainly have to be considered were we dealing with authentic (spontaneous) speech, where glottalisation is intimately involved in cases of hesitation, faltering, false start, and the like.)

Since glottalisation often *substitutes* for a [-VOICE] stop or a [ch]-closure, it is not surprising that it may also occur as an accompaniment to the same, partly perhaps through a kind of contamination by frequent association, and partly (also a 'perhaps'!) because of the speaker's awareness that there are circumstances in which a burst may be inaudible (as may be the case with the brief bursts before [-VOICE] fricatives), and thus from a desire to reinforce the perception of a stop through the production of an effect (laryngealisation of the preceding vowel) associated with [-VOICE] stops elsewhere (figures 2.30, 2.31, 2.32). This kind of glottalisation will be referred to in the conventional way as STOP REINFORCEMENT. Thus we have a basic division in the functions of glottalisation as between cases associated in some way with stops, to which we may refer collectively as STOP MARKING, and cases associated with juncture, giving us the classification-scheme shown in figure 2.33.

Unfortunately, while we may be able to analyse particular cases of STOP MARKING in the terms just outlined when we have access to the linguistic representation of the utterance concerned, when we have only the acoustic record it is not always possible to work out with confidence exactly what analysis applies; the reasons for this will be given a little further on. For recognition-purposes, therefore, our policy has in certain cases to be a rather agnostic one, as far as the earliest stages of processing are concerned.

### 2.7.1 Clear Cases of Stop Substitution

I count as clear cases of stop substitution all cases without a period of either silence or merely residual voice between the phone 'preceding' the notional stop and that 'following' it. In all such cases, it is impossible not to recognise that

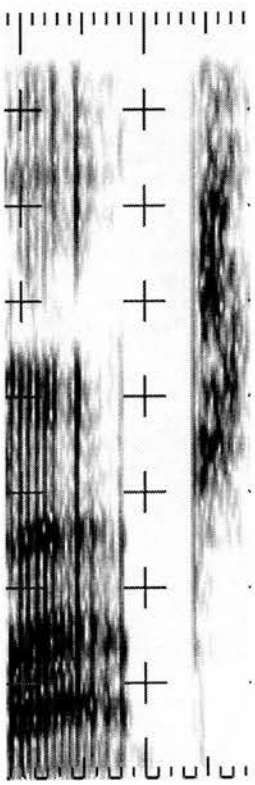


**Figure 2.30.** [a GLpc pb s] from utterance of "collapsed" (sc150)

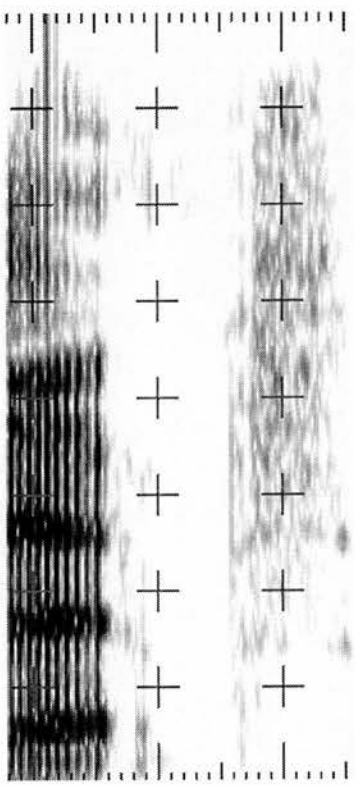


**Figure 2.31.** [e GLpc pb th] from utterance of "depth" (sc111)

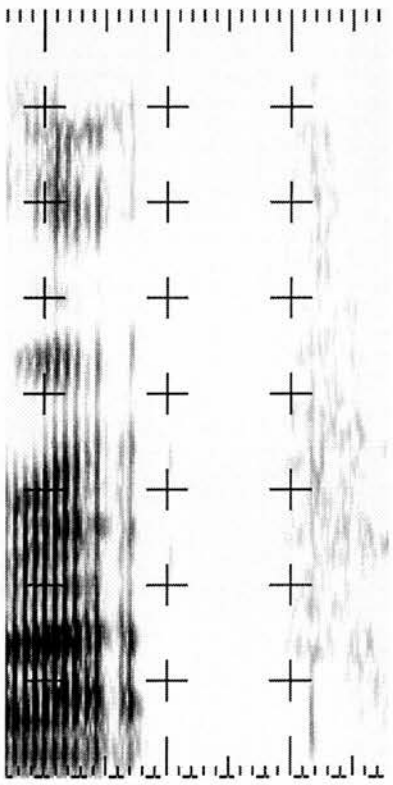


**Figure 2.32.** [a GLpc f] from utterance of "von Trapp family" (sc122)

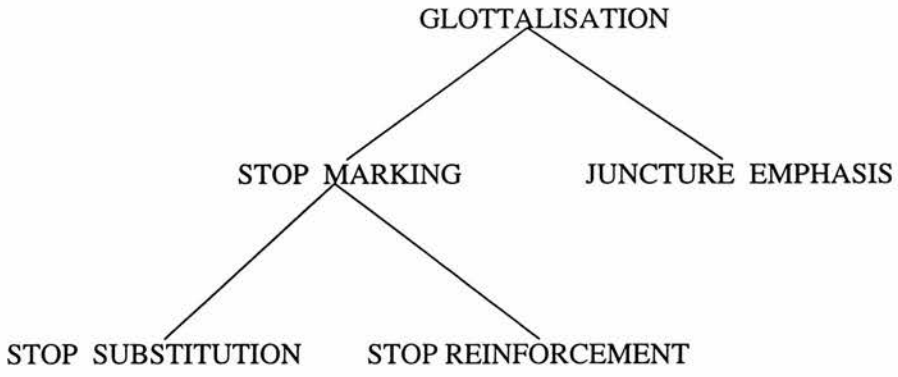



**Figure 2.33.** Forms/Functions of Glottalisation

| utterance | manual transcription | expanded to |
|---|---|---|
| it was | [ i GLw1 w2 ax z ] | [...i_GLwA GL_wB w_axA ...] |
| it felt | [ i GLf e dl t ] | [...i_GLfA GL_fB f_eA ...] |
| don't shout | [ d ou n GLsh au t] | [...n_GLshA GL_shB sh_aD1A ...] |

**Table 2.15.** Automatic Conversion of Glottalisation Labels

we are dealing with a non-segmental phenomenon at the acoustic level, even though it is one which arises from the speaker's desire to indicate a difference that would be represented by a discrete symbol in a phonemic or orthographic representation. The strategy adopted for handling this is to create a zero-duration acoustic 'segment' at the boundary between the 'preceding' and 'following' phone, and to allow its presence to cause the annotation of the offset of the 'preceding' phone to indicate that it is affected by a stop substitution, and similarly with the onset of the 'following' phone. In active recognition, a correct transcription in subphonic terms (a correct AS-transcription) can thus give rise to the writing-in of a discrete stop-label in the representation produced at subsequent levels.

No attempt is made (prior to and including the stage of AP-transcription) to identify the PLACE of the substituted stop, decisions about its PLACE being postponed to later stages of processing ([t] is the most commonly substituted stop, but it is not, of course, the only one). 'GL' is prefixed to the 'following' phone in manual labelling, to indicate the presence of a preceding glottalisation and subsequent automatic expansions take place as illustrated in table 2.15. In this way, the fact that the glottalisation is, so to speak, superimposed on the sequence of phones, rather than occupying a distinct section of the acoustic record in its own right, ceases to present any difficulty.

In many of these cases there does not seem to be any gesture toward the POA of the substituted stop, though I suppose it is not impossible — certainly in the case of [t] substituted before [w], and [r] — that the acoustic effect of such a gesture may in some cases be simply submerged by the more drastic effects of significant lip-rounding and jaw-movement. Ideally, the onset of the 'following' phone in these cases would be marked not only as following or (for [+VOICE] phones) being involved in a laryngealisation , but also to indicate the identity of

the 'preceding' phone; but shortages of tokens forced a compromise here.

It is worth noting that in some of these cases the PLACE of the substituted stop is recoverable on a priori grounds; with preceding nasals, e.g., the homorganic nasal-stop rule will deliver the PLACE, no matter how good or bad the acoustic evidence may be, and it seems more sensible therefore to rely on the rule than on the acoustic record.

### 2.7.2 Stop Marking of Doubtful Species

In all the cases to be discussed under this heading, a period of merely residual or strangulated voice and/or silence is evident in the spectrographic record, with the preceding vowel, [dl], or nasal clearly laryngealised in its later stages. It is difficult to know, from the acoustic record alone, whether only a single stop is involved, or whether there may not be a pair of stops, with argument then possible over whether the first stop is unreleased and glottally reinforced, or wholly substituted. (Though a very brief closure-period would argue in favour of a single stop glottally reinforced, duration is not explicitly taken into account in processing to derive the AS- or AP-transcription, so that duration of the closure is not brought into consideration up to that point.) In cases where two stops are definitely involved, as far as the linguistic content is concerned, as in productions of *credit card* or *that part* with glottalisation, the formant pattern in the offset of the laryngealised sonorant may or may may not show unequivocally whether the POA approached is that for the first or that for the second stop; depending on the relative timing of glottal and supraglottal gestures, phonation may cease altogether before the formant pattern transition has proceeded far enough to be able to tell, or a POA may be approached that could represent a compromise between the POA's of the individual stops. There are cases in the data where the transition is *clearly* to the POA of the *second* stop, but there are many cases where the spectrographic evidence is not unequivocal.

Before providing examples of how the glottalisations of this section are handled overall, it is necessary to say something about labelling-policy where pairs of consecutive stops are involved (at the linguistic level at least). Representing the first stop as x and the second as y, if the formant pattern in the offset of

the preceding phone showed a movement toward the POA of [x], the stop closure was given the complex label 'GLxyc' (hence [GLtkc], [GLptc], etc.), but if the formant pattern revealed rather a movement directly toward the POA of [y], the stop closure was given the label 'GLyc' (as e.g. for at the end of the word 'credit' in sentence 069 where the label [GLkc] was used – figure 2.34). It was not always possible to be confident about the decision, of course, but it seemed worthwhile attempting to implement the policy for the sake of the cases that were clear, and because there seemed to be some hope of being able to derive some duration-statistics from the labelling. However, none of the specific information in these stop-closure labels is made use of in subsequent processing down to the stage of AP-transcription. The specific closure-labels are automatically generalised, as described for normal stop-closures in section 2.5.2, after the 'GL' prefix has been transferred to the offset of the preceding phone. With normal stop-closures (those not involving any glottalisation-process) the interpretation of a segment like [NVSAc] in AP-transcription is dependent upon the subphones preceding and following it, and effected at the stage of AP-transcription; for example, if the subphone preceding the closure is marked as preceding a [t], while the subphone that follows it is a (contextually annotated) [kb], the [NVSAc] segment is converted to [tkc]. However, given the unreliability of PLACE information in glottalised offsets, this procedure would be liable to lead to errors. No interpretation is attempted, therefore, at such an early stage. All that is looked for from the initial automatic classification is an indication that here a closure occurs following on from a glottalisation connected with it. Hence it remains an open question whether we have one stop or two, and an open question, too, what the PLACE of the stop, or PLACEs of the stops, may be.

In all the cases under this heading, then, the 'GL' prefix is attached to the stop-closure label, so that for example for the spectrogram shown as figure 2.30 we get the labelling [kc kb ax cl a GLpc pb s]. The offset of the preceding phone gets automatically marked as glottalised and as preceding a stop-closure, but with no commitment either to the number of stops involved or to the PLACE or PLACEs of the stop or stops. In the subsequent processing of the stop-closure itself, the 'GL' prefix is ignored, and the closure is generalised and treated in the same way as any other stop-closure, in the way described in section 2.5.2 above on stops.

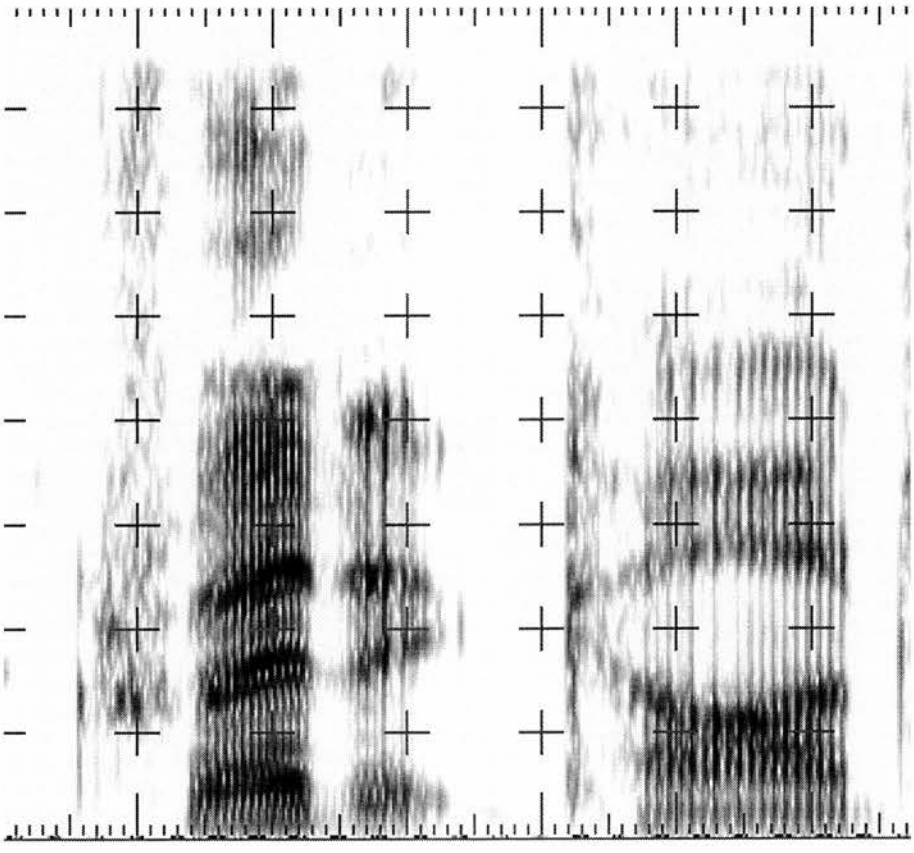


**Figure 2.34.** [kc krb r2 e dnc ir GLkc kb aa d] ("credit card")

| utterance | manual labelling | automatic conversion |
|---|---|---|
| collapsed | [kc kb ax cl a GLpc pb s utc Tb] | [...a_GLscA NVSAc pb_sA ...] |
| credit card | [kc krb r2 e dnc i GLkc kb aa dc db] | [...i_GLscA NVSAc kb1_aaA ...] |
| work for Tom | [w1 w2 lax GLkc f ax tc tb o m] | [...lax_GLscA NVSAc GL_fB ...] |

**Table 2.16.** Automatic Conversion of Glottalisation Labels

In processing the stop-release, too, (if one is present), the earlier glottalisation is ignored, stop-releases being made sensitive only to their right context. It follows that the AS-transcription will signal the presence of the glottal stop reinforcement simply and solely through the glottalisation of the offset of the preceding phone, at least if the stop has a release. In the case of stops with no release, there is room for debate about the best way to annotate the following phone's left context, even though this is one context of occurrence where we can safely assume that we have only a single stop (as far as phonetic realisation is concerned), the second of any pair of stops involved in a glottalisation being virtually certain to be released. The reason for the uncertainty is that a closure-period can be the result either of an oral occlusion (which would cause the onset of the following phone to carry some PLACE information) or of the glottal constriction (in which case the onset of the following phone might carry little or no PLACE information, depending on whether any complete oral closure accompanied the glottal stop). In some tokens, annotating for the bare glottalisation would be most appropriate, in others annotation for the PLACE of the stop. Since we are allowed only a single policy, it was decided to annotate for bare glottalisation, partly just to allow data-sharing with some of the cases discussed in the previous section (both in stop-substitutions, and in the cases currently under discussion, we might for example get an [f] as the 'following' phone — in both cases its onset would become a [GL_fB]). Some examples are given in table 2.16.

The postponement of decisions makes sense in a number of ways. Quite apart from the riskiness of attempting the decisions earlier, on the basis of acoustic evidence alone, there are almost certainly structural regularities of a phonological kind governing the distribution of stop-reinforcement and substitution, which can be gleaned from across the whole of the data for a given speaker, and perhaps even

for a particular accent, so liberating us from total servitude to data-quotas (that is, to the need for amounts of data for each case we wish to model that are large enough for statistical modelling). Even if only as heuristics, rules based on such regularities may be able to provide useful probabilities for working systematically through the possible interpretations of the evidence. For GSW, it can be said that glottalisation is possible before a single stop-closure (i.e. as mere reinforcement), and even likely when the stop is released into a [-VOICE] fricative, and that glottalisation is much less probable before a single stop that is released into a following vowel, so that if an AP-transcription indicates a glottalisation in such a context there must be a greater than even probability of there being a pair of stops at the linguistic level.

The agnostic strategy also has some limitations, needless to say. In a single distribution for, e.g., all offset-subphones of some glottalised vowel before stop-closures of all types, there *will* be tokens with clear trends toward an alveolar, velar, or labial POA, as well as nondescript tokens, the common feature of all tokens being their being affected by glottalisation. Given a model trained from such data using the assumption of Normality, a token with clear formant pattern trend toward a definite POA might possibly score more highly as its non-glottalised counterpart, with consequent loss of the information that glottalisation is present. This is perhaps a case where mixture-modelling recommends itself, with modes for each of the three POA's.

### 2.7.3 Juncture Emphasis

From a processing point of view, juncture emphasis ought to be far simpler to handle than stop marking, but in fact it is impossible to handle it satisfactorily with limited data just because examples are so widely scattered across different phonetic contexts, making statistical modelling difficult.

With the Normality assumption, it is necessary to introduce a dichotomy, with cases of complete, sustained glottal closure on the one hand (these will be referred to as 'full juncture glottalisation') and cases with no such fast closure on the other ('weak juncture glottalisation').

**Full Juncture Glottalisation**

The glottal stricture here is fast and sustained, even if only momentarily. Between two vowels, the articulatory scheduling *seems* from the data (but the data is not plentiful) to involve putting a halt to the phonation involved in the first vowel before making any of the supraglottal changes required for production of the second vowel, and to relax the glottal stricture only when these movements have been completed. In labelling, the period of closure is labelled 'GL', and the annotation of the offset of the preceding and onset of the following vowel follows automatically after the normal pattern ([GL] is in this respect like any normal segment). Unfortunately, there is insufficient data for good modelling of either offset of the first or onset of the second vowel, for any vowel in the data.

Where the glottal closure is used to produce a silence between a consonant and a vowel, the offset of the consonant is likely to be rather *sui generis* — e.g. quite unlike pre-pausal offsets, because of the abrupt termination — but dedicated statistical modelling proved impossible because of the shortage of data for any particular pair of consonant and vowel. The policy adopted — not a very satisfactory one, but the best that could be found — was to assume that there would likely be some tendency for the consonant to show articulatory anticipation for the vowel that would follow the oral closure, and so to lump the case with others involving the same consonant and vowel but without any intervening glottal closure or indeed any glottalisation at all. In order to get the required annotation across the intervening glottal closure, the closure is given a label which also indicates the identity of the vowel that follows, e.g. 'GSii' when the following vowel is an [ii]. The vowel itself, as with the two-vowel case discussed in the previous paragraph, has its onset annotated automatically as following on from a glottal closure. Hence a production of "...has easily ..." with full juncture glottalisation could be labelled as [h ax z GSii ii z ir l ii] just in order to secure the desired contextual annotations. Sequence-constraints then have to allow all offsets of consonant X with vowel Y as right context to be capable of being followed also by [GSY], because of the lumping referred to earlier in the paragraph.

### Weak Juncture Glottalisation

In the cases under this heading, the glottal stricture is not forceful enough to provide a clear closure period. When the stricture occurs between vowels, 'GL' is prefixed to the label for the second vowel, and — again because of extreme data-shortages — the offset of the first vowel is automatically annotated as preceding a glottal stricture, and the onset of the following vowel is annotated analogously. (In some cases, particularly where the glottalisation is very weak, it would make as much sense to annotate the vowels simply with respect to the vocalic successor or predecessor, but many non-glottalised vowel-vowel transitions are also poorly represented in the data; certainly there is no hope of modelling specific glottalised cases aiming to capture both glottalisation *and* vowel identities.)

When there is weak juncture glottalisation between a consonant and a vowel (figure 2.35 and 2.36), I prefix 'GL' to the label for the vowel, but as with the corresponding full juncture glottalisations, the consonant-offset comes to be automatically annotated as preceding the vowel that follows it, ignoring the glottalisation (glottalised and non-glottalised cases are lumped together), while the vowel-onset is automatically annotated as beginning from a glottal stricture (ignoring the identity of the consonant).

A crude attempt was made to bolster cases of vowel-onsets following glottal stricture by lumping vowels with such onsets automatically with sentence-initial vowels, some of which exhibit similar acoustic traits in their onsets. It was intended in a later implementation to screen the sentence-initial vowels and explicitly label just those that do have such traits, and to lump only with these, but at the time of writing this has still not been done.

There is of course no substitute for real data, but an idea considered for trying to boost the amount of data for offsets of vowels before juncture-emphasising glottal stricture was to take data from identical vowels following such stricture, and to reverse the sign of just those values representing trend-information (see Chapter 3, section 6), leaving everything else unchanged. This has not been tried to date, however.

Glottalised vowel-offsets and onsets are another case where the Normality assumption is less than ideal, since one may find widely scattered pitch pulses

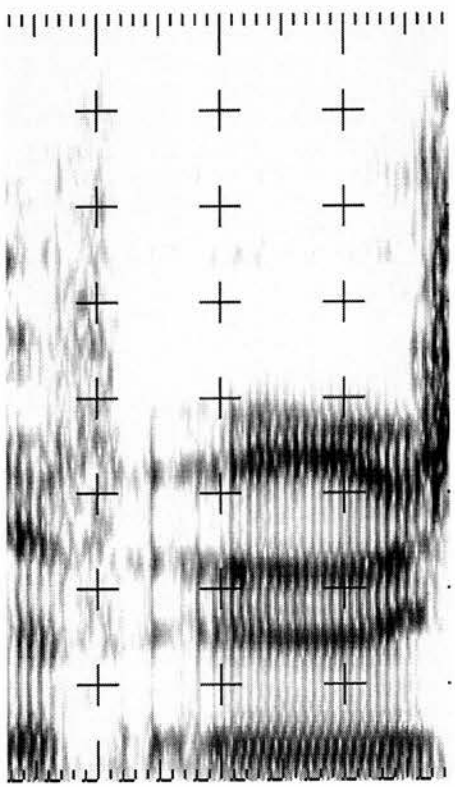


**Figure 2.35.** [ax v GLuuf z] from utterance of "have oozed" (sc122)

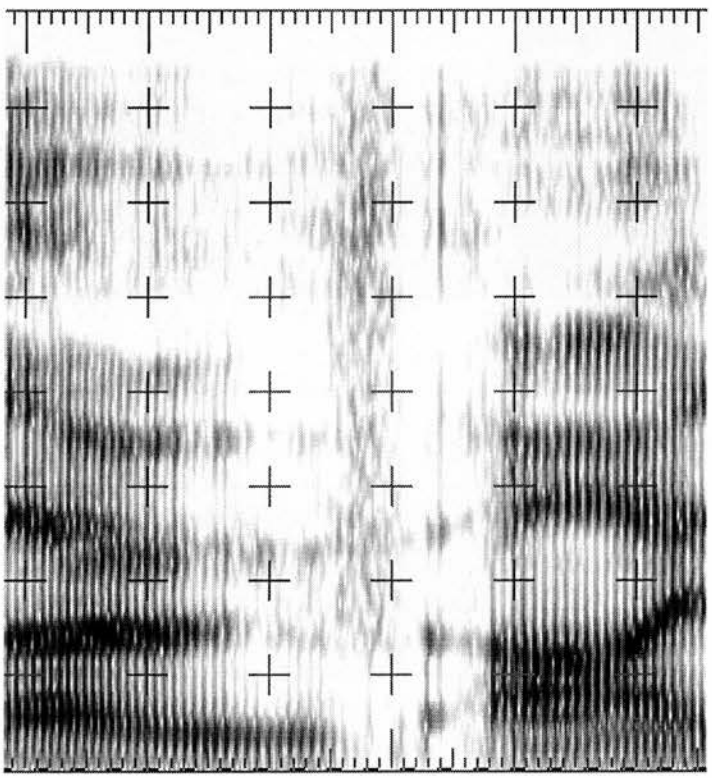


**Figure 2.36.** [ou PVdh GLaaD1 aa2i iD3] from utterance of "loathe eiderdowns" (sc122)

and intervening stretches of diffuse low-amplitude noise or silence, and at least a bimodal distribution would seem necessary to do justice to such data.

# Chapter 3

# Representing the Short-term Spectrum

## 3.1 Introduction

The main purpose of this chapter is to define and explain the form chosen in this work for representing the short-term characteristics of speech. There is little that is original in the representation used, and I have therefore concentrated my efforts on trying to write the chapter in a way which will make its content accessible to readers with little or no knowledge of signal processing, other than a basic understanding of the principles underlying traditional spectral analysis. The account of cepstral analysis (based on treatments by Noll (Noll 1964), Schafer and Rabiner (Schafer & Rabiner 1975), and O'Shaugnessey (O'Shaugnessey 1987)) is given in this spirit.

In section 3.4.3 I discuss the role of critical bands in human auditory perception, and their significance for machine recognition. I argue against a too slavish adoption of critical bands in the application of frequency-transformations, while acknowledging the importance of such transformations and the significance of the Bark and Mel scales in this connection. In section 3.5 I explain the reasons for introducing classes to model boundary-regions between phones, and provide examples of some of these 'border-straddling' classes (a fuller treatment is reserved to an appendix). Finally in section 3.6 an account is given of additional features

designed to capture what is happening over the slightly longer-term (longer, that is, than that considered in the basic representation), together with some discussion of the advantages and disadvantages of using these features.

## 3.2 The Short-term View

The discussion in Chapter 2 operated chiefly at a phonetic level, in terms of articulatory movements, phones, subphonic elements, and the like. The initial input to a recognition-system, however, typically consists of a bare series of numbers representing successive amplitude values of a sampled speech waveform (or of its electrical analogue created via a microphone). While the behaviour of the numbers in this series is determined by the underlying sequence of phonetic events, some form of analysis is required to make this relationship perspicuous, but as the phonetic events themselves are unknown at the outset, we are not in a position to be able to take sub-series corresponding to particular phones and apply an analysis to them. We are obliged rather to work automaton-like, taking some fixed-length subseries of the numbers at a time, carrying out an analysis, and repeating the procedure again and again from the beginning of the utterance to the end. At each point we view the series of numbers through a window which comprises just a small subseries from the total, shutting out anything earlier or later; we will refer to one such subseries comprised by the analysis-window as falling within, or more loosely as constituting, a *frame*. The numbers we get as a *result* of each analysis will come to stand in place of the original frame of data, hence to 'represent' it, and it is convenient to give a fixed order to the numbers arrived at, to arrange them, that is, in the form of a vector. These ordered lists of numbers, or vectors, which stand in place of the original signal, will be referred to henceforth as representation-vectors or pattern-vectors (or sometimes as just vectors), and constitute our representation of the short-term spectrum. (It is sufficient for present purposes to think of 'vector' as *meaning* no more than an ordered list of numbers.)

Because the speech-waveform is continually changing, we need to perform our analyses at very frequent intervals if we are not to miss anything of interest. Moreover, since change is continuous, we must take a fairly short window each

time as material for the analysis; we will in fact be assuming, in defiance of the exact reality, that the signal is unchanging or *stationary* within the confines of the analysis-window, and the shorter the window, the less objectionable this assumption would appear to be. On the other hand, it will turn out — for reasons to be touched on below — that too short a window will not provide suitable material for analysis, so that a compromise has to be reached.

It is common practice to use a window of about 20 ms, and to advance the window each time by some fraction of this, say a quarter or a half, so that successive analyses are overlapping. It is normal practice, too, to apply a window-function to the waveform numbers within a window, to give greater emphasis to the numbers at the centre of the window and progressively less emphasis to numbers further out toward the edges; this is simply a multiplication by a factor which reaches 1 around the centre of the window and tapers off smoothly toward zero at the window's outer limits. The use of a windowing function serves two purposes: firstly, it helps to make the stationarity assumption a little less unrealistic for a long window, by giving selectively greater emphasis to a central portion within it, and secondly it reduces the effect on analysis-results of advances of the window-position that would otherwise cause major discontinuities, such as those which involve the exclusion or inclusion of a pitch-pulse (the problem of *edge-effects*). Edge-effects arising from inclusion or exclusion of a pitch-pulse as the analysis-window moves along are also minimised by use of a longer window, for speakers with normal or high fundamental frequencies at least; the more pitch-periods comprised by a single window, the less significant the dropping-out or the taking-in of a single pitch-pulse will be (O'Shaugnessey 1987).

By working through the utterance in this way we are able to track the changing signal, window by window or frame by frame, without excessive smearing of information. I turn now to consider some of the features that are desirable in the representation-vectors so produced. These vectors are to be used as the basic currency of the next stages of the recognition-process.

## 3.3 Features Desirable in a Representation Vector

A representation of the short-term characteristics of the speech-signal should be economical, trustworthy, and above all possessed of discriminant power; for statistical recognition-systems, it may also be an advantage if its components are not strongly correlated.

If two representations were available that scored equally well with respect to criteria other than economy, but one involved several hundred components and the other only a dozen, the latter would be greatly preferable in any computational context, and particularly in systems using traditional (non-connectionist) statistical pattern-classification, where *dimensionality* — the number of elements in the representation-vector — has a crucial impact on classifier-performance given finite amounts of training-data. More will be said on this (and on the relevance of correlation between components) in Chapter 4.

It is clearly also desirable that the process of deriving a particular representation should not be prone to serious or pervasive error. While formant-based description of vowels in particular is extremely economical, the automatic tracking of formants is fairly error-prone, which makes formant representations less than ideal for automatic recognition.

The importance of discriminant power hardly needs argument. It will not always be the case that short-term spectral information will point unambiguously to the underlying phonetic event, but it is obvious that all information that will *contribute* to the identification should be captured if at all possible. It may well prove, too, that *separating* certain features of the speech-signal, such as fundamental frequency, power, duration and spectral envelope, may make phonetic identification easier.

## 3.4 Mel Cepstral Representations

Representations based on mel cepstral coefficients score either well or reasonably well with respect to all the criteria discussed, as will be shown in what follows,

and for this reason are in widespread use in ASR. In this section I first present cepstral analysis as a means of extracting the most salient features of the spectrum in a highly economical form. It is customary to present cepstral processing as a means of separating out the contributions of source- and filter- components to the output speech spectrum, and I shall go on to consider how accurate such a picture can be.

### 3.4.1 Extraction of the Most Salient Features of the Spectrum

If we take an arbitrary log magnitude spectrum, and treat it as if it were itself a time-domain signal, i.e. a waveform, we may allow ourselves the liberty of thinking about analysing it into its component 'frequencies' Fourier-style, as we do with a true waveform, thus establishing the amount of energy at each different 'frequency'. Intuitively, we would then expect to find the more slowly-varying features of the spectrum being registered at lower 'frequencies', and progressively more rapidly-varying features by 'frequencies' higher and higher up the 'frequency'-range. By applying an inverse DFT (IDFT) to the (log magnitude) spectrum to produce the cepstrum, this is in fact what we achieve, except of course that we cannot use the phrase 'frequency-components' as I have done in quotes here: the 'frequencies' are actually 'quefrencies', and in a display of the cepstrum the x-axis or quefrency-axis is measured in units of time.

The derivation of the cepstrum from the log magnitude spectrum can be presented in extremely simple terms. Using a discrete cosine transform (DCT) in place of the IDFT (the IDFT uses both sines and cosines, the DCT only cosines), the formula for the derivation is as follows:

$$cc_i = \sum_{k=1}^{N} X_k \cos\left[i(k-0.5)\frac{\pi}{N}\right], \quad i = 1, 2, \ldots, M,$$

where $cc_i$ is the $i'th$ cepstral coefficient, and $X_k, k = 1, 2, \ldots, N$ are the log energies of the Fourier spectrum.

Consider the term $\cos[i(k-0.5)\frac{\pi}{N}]$ which defines the *basis functions* of the transform. The first 8 basis functions are plotted in figure 3.1 and figure 3.2,

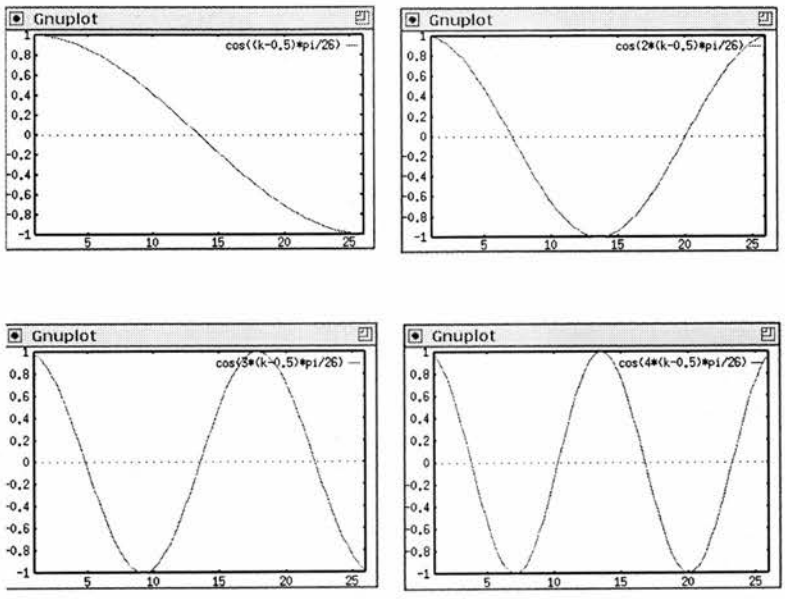



**Figure 3.1.** First Four Basis Functions of DCT, showing transform coefficients on the y-axis for values of x representing ordinal points in the series of FFT spectral coefficients

with $N$ set to 26, the 26 steps being ranged along the x-axis of each plot, and transform coefficients shown along the y-axis.[1] The value of 26 for $N$ will be explained in due course.

Each cepstral coefficient is obtained by multiplying the (log magnitude) spectral value by the corresponding transform coefficient, i.e. by values of between 1 and -1 depending on the alignment between the spectral coefficient and the particular basis function, and then summing the products. The first cepstral coefficient can thus easily be seen to represent the bias toward low or high frequency energy in the spectrum, with 'low' meaning below the half-way point in a linear spectral frequency-range, and 'high' above that point.[2] Given a linear frequency-scale

[1] I have excluded the so-called 0'th coefficient from these plots. This coefficient is a constant with the value of 1, and is sometimes used as a measure of total energy.

[2] One should try to imagine the plot of each basis function in turn being overlaid on the spectrum of selected phones, say of the vowel [oo] and of [s]. As the plot of the first of the basis functions shows, all the spectral coefficients in the lower half of the frequency-range get multiplied by positive values, while all those in the upper half get multiplied by negative values.

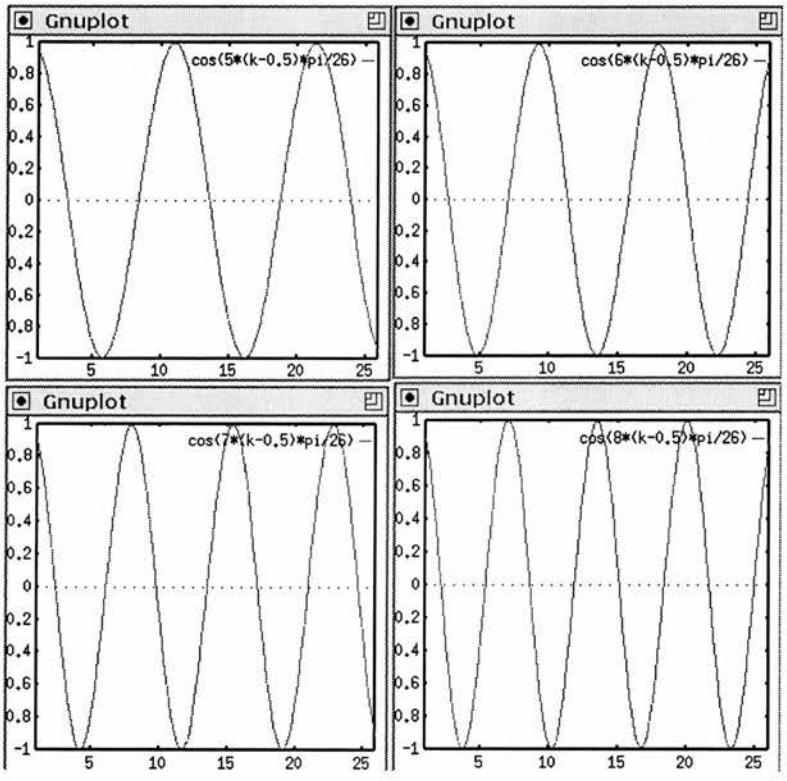



**Figure 3.2.** Fifth to Eighth Basis Functions of DCT, axes as in previous figure

extending from 0 to 6 kHz, vowels would get relatively large values for the first cepstral coefficient, while a phone like [s] with virtually all its energy above the midway-point would get a highly negative first coefficient, all its higher spectral coefficients being multiplied by negative values.

It is fairly easy — *for a definite frequency-scaling, and a definite speaker* — to work out what sorts of values particular phones are likely to get for the second, third, fourth and fifth coefficients, as a perusal of the basis functions will show. But it becomes increasingly true as one ascends through the basis functions — and particularly if using a frequency-scale that compresses certain frequency-regions — that 'accentuations' and 'diminutions' come to cancel each other out more and more, so that it becomes harder to arrive at conclusions merely intuitively. That is to say, by the time one reaches the 7th or 8th basis function, with several periods of the function comprised within the frequency-range, the chances are that high products arising from energy within a region multiplied by positive factors will be neutralised by low products arising from energy within regions multiplied by negative factors. The higher-frequency cosines of basis functions beyond the first dozen or so in any case become less and less significant, with only one qualification. The one or two basis functions which most nearly coincide with the 'voicing-ripple' evident in the log spectrum for a voiced phone will register a massive increase over those that precede and follow them (the positive phases of their cycle coinciding with the peaks of the voicing-ripple), and hence give rise to a distinctive spike or impulse in the cepstral display at a quefrency equal to the reciprocal of F0. This apart, it is generally held to be true that just the first dozen or so coefficients are sufficient to provide good discrimination between spectra.

---

Since the cepstral coefficient is then obtained by summing the products of these multiplications, it should be clear that sounds with most energy in the lower half of the frequency-range will get relatively high values for this first coefficient, while those with most of their energy in the upper half of the frequency-range will get relatively low (negative) values.

### 3.4.2 Separation of Contributions from Source and Filter

In speech, a sound-source — voice, frication-noise, or both — is modified in accordance with the shape-dependent (and of course time-varying) filtering characteristics of the vocal tract, with energies at certain frequencies in the source being amplified and others attenuated. For certain purposes it is advantageous to isolate the characteristics of the speech spectrum that chiefly result from features of the source from those that chiefly reflect the vocal tract frequency-response. Cepstral processing is often presented as a means of effecting such a separation.

For vowels and sonorant consonants, the source-spectrum consists of evenly-spaced harmonics at whole-number multiples of the fundamental frequency. The spacing of harmonics, then, varies as a function of F0. The glottal source-spectrum has a negative slope (tilt, roll-off) with ascending frequency of about -12 dB per octave *on average* but the slope varies significantly as a function of voice-intensity and phonation-type (Fant 1995). The slope is determined chiefly by the way airflow changes within a phonatory cycle (Clark & Yallop 1990): as voice intensity increases, there is a corresponding decrease in the open quotient (the ratio of the duration of the open glottal phase to that of the closed glottal phase within a phonatory cycle) (Veeneman & BeMent 1985), and this is reflected in greater gradualness in the roll-off in the source-spectrum, with more energy surviving at higher harmonics (Clark & Yallop 1990). Fant states (*op. cit.*) that a decrease in OQ, as in a pressed voice, promotes the level of the second harmonic at the expense of the fundamental. In breathy voice, on the other hand, with a slow closing action within the phonatory cycle, and minimal closed phase, the roll-off in the source-spectrum tends to be steep, with little energy surviving at higher harmonics (Clark & Yallop 1990). Fant speaks of there being a relative boost to the fundamental in breathy phonation (*op. cit.*).

Now while the frequency-response of the vocal tract is a matter of its own shape at any instant, it should nevertheless be clear that variation in the spacing of glottal harmonics, and in the relative strengths of individual harmonics (or more generally, in the spectral tilt of the glottal source), will play a significant role in determining the overall shape of the output speech spectrum, and that it is therefore an exaggeration to speak of cepstral analysis being able to fully

isolate the vocal tract frequency-response (I take it that this is the reason for Rabiner and Schafer's use of the word "approximate" in respect of the deconvolving of source and vocal tract impulse response in cepstral processing (Schafer & Rabiner 1975)). The degree of resonance at any resonant frequency, for example, will depend on the degree of coincidence in frequency with the nearest harmonic(s), and on the energy in the latter. Again, Fant shows (*op. cit.*) that differences in voice-effort produce non-linear shifts in the distribution of energy in the output speech spectrum, with higher-frequency energies gaining more than lower ones (for sustained vowels, a 10 dB increase in the level of F1 was found to be accompanied by a 4 dB increase in the level of the fundamental, a 14 dB increase for F2, 16 dB for F3 and 14 dB for F4; for natural connected speech the shifts were not so dramatic, but still significant). Thus it is strictly speaking a mistake to say, for example, that cepstral coefficients "do not depend upon the frame energy" (Pirani 1990). What *is* true is that if we took two spectra, with one equal to some constant multiple of the other, the cepstral coefficients for the two spectra would be identical (if we ignore any 0'th coefficient). However, since differences in voice-intensity reflect themselves in the speech spectrum in the way indicated, cepstral representations will not abstract completely from such differences. From the point of view of context-sensitive phonetic classification it is in fact fortunate that this is so, since it means that use of a cepstral representation does not cause us to lose information that can be important in identifying such things as regions of voicing-switch, as in vowel-offsets before [-VOICE] fricatives, where the final pitch-periods of the vowel are likely to be marked by increasing breathiness and loss of amplitude.

Cepstral representations score highly in respect of the criteria detailed in section 3.3 above. That they are very economical (relative to the Fourier spectrum itself) has already been shown. Given reasonable choice of window-size, their derivation is not prone to error. The coefficients are much less highly correlated than, e.g., filter-bank energy coefficients (Pirani 1990), though their lack of correlation should not be exaggerated (the effects of assuming complete statistical independence of the coefficients will be considered in Chapter 5 (5.5)). As far as the all-important feature of discriminant power is concerned, it should be clear that the ability to represent the essential characteristics of the speech spectrum,

in abstraction from what is mostly superficial detail, offers significant advantages. Cepstral representations of one form or another are used in most current ASR systems, itself a testimony to their ability to deliver effective phonetic discrimination. When used in conjunction with a non-linear frequency-scale they have been found more effective still, and it is to the issue of frequency-transformations that I now turn.

### 3.4.3 Frequency-Warping and Critical Bands

Cepstral coefficients may be derived directly from the log magnitude spectrum as described, but it is common practice to modify the spectrum before the derivation, the purpose of this normally being explained in terms of choosing a frequency-resolution that conforms more closely to that of which the human ear is capable.

The Fourier spectrum represents the energies within each of a number of bands centred on frequencies linearly spaced across the frequency-range, but human hearing works along rather different lines from this. Psychophysical investigations have established the existence of *critical bands* in human auditory perception. Two concepts that are important for an understanding of critical bands are those of *threshold* and *masker*. The threshold of a stimulus (a pure tone, for example) is the strength at which it is *just* audible; a masker is a tone or sound which is able to obscure perception of the stimulus. The following account from Scharf introduces the most important facts about the critical band:

> As a purely empirical phenomenon, the critical band is that bandwidth at which subjective responses rather abruptly change ... . Thus the loudness of a band of noise at a constant sound pressure remains constant as the bandwidth increases up to the critical band; then loudness begins to increase. In another type of experiment, the threshold of a narrow band of noise lying between two masking tones remains constant as the frequency separation between the tones increases until the critical band is reached; then the threshold of the noise drops precipitously. In these and other experiments, the critical band requires manipulation of bandwidth. Measured in this manner, the critical band turns out to be remarkably alike in many types of experiment.

(Scharf 1972)

The width of the critical bands varies as a function of their centre-frequencies, as indicated in table 3.1 from Zwicker (Zwicker 1961), which clearly reveals the differences in sensitivity at different points of the frequency-range.

Zwicker and Terhardt (Zwicker & Terhardt 1980) give analytical expressions for converting a linear frequency-scale to a critical band scale. One critical band is equal to one interval of the bark scale, and 1 bark is very nearly equivalent to 100 mels. A transformation of the linear frequency scale to a bark or mel scale can be approximated (Bengio 1996) as follows (where B(.) is the transformation function and f is the frequency in Hz):

$$B(f) = 0.01f \qquad f < 500Hz \tag{3.1}$$

$$B(f) = 0.007f \qquad 500Hz < f < 1220Hz \tag{3.2}$$

$$B(f) = (6 \ln f) - 32.6 \qquad f \geq 1220Hz \tag{3.3}$$

Several points need to be made about the published accounts of the critical bands, and a distinction needs to be drawn between critical bands established by direct experiment (the *empirical* critical bands, as Scharf refers to them) and schemes designed to simulate the supposed frequency-analysis function of the ear. First as regards the figures given in table 3.1: Zwicker explicitly states (*op. cit.*) that while the *bandwidths* of the critical bands are relatively fixed, their *position* on the frequency-scale is more variable, and their position can be *altered continuously*, "perhaps by the ear itself". Scharf states that while there is generally close agreement between published accounts of measurements of critical bands, the figures are probably reliable only within about + or - 15% because of variability between subjects and indeed even in respect of a single subject. Scharf also states that in many threshold experiments critical bandwidths increased for shorter-duration sounds (below about 100 ms), the latter statement in particular having obvious relevance to any involvement of critical bands in speech perception.

As to the distinction between the proven existence of the empirical critical band and theory about the ear's supposed ability to do a frequency-analysis on

| Critical Bands (Zwicker 1961) | | | |
|---|---|---|---|
| Number | Centre Frequency (Hz) | Cut-off Frequency (Hz) | Bandwidth (Hz) |
| 1 | 50 | 100 | 80 |
| 2 | 150 | 200 | 100 |
| 3 | 250 | 300 | 100 |
| 4 | 350 | 400 | 100 |
| 5 | 450 | 510 | 110 |
| 6 | 570 | 630 | 120 |
| 7 | 700 | 770 | 140 |
| 8 | 840 | 920 | 150 |
| 9 | 1000 | 1080 | 160 |
| 10 | 1170 | 1270 | 190 |
| 11 | 1370 | 1480 | 210 |
| 12 | 1600 | 1720 | 240 |
| 13 | 1850 | 2000 | 280 |
| 14 | 2150 | 2320 | 320 |
| 15 | 2500 | 2700 | 380 |
| 16 | 2900 | 3150 | 450 |
| 17 | 3400 | 3700 | 550 |
| 18 | 4000 | 4400 | 700 |
| 19 | 4800 | 5300 | 900 |
| 20 | 5800 | 6400 | 1100 |
| 21 | 7000 | 7700 | 1300 |
| 22 | 8500 | 9500 | 1800 |
| 23 | 10500 | 12000 | 2500 |
| 24 | 13500 | 15500 | 3500 |

**Table 3.1.** Critical Band Specifications

speech-sounds, Scharf is worth quoting again in this connection, though thirty years of research have passed since he wrote:

> The role, if any, of the critical band in the perception of speech remains obscure even though the initial measurements of the critical ratios [signal to noise ratios used to determine critical bands] were closely related to search for a meaningful analysis of the speech spectrum. ... French and Steinberg (French & J.C.Steinberg 1947), measuring the intelligibility of speech passed through low and high cut-off filters, found that 20 adjacent frequency-bands contribute about equally when each approximates a critical band. *The width of the 20 bands does not grow as rapidly with frequency above 1500 Hz as does the critical bandwidth* (my emphasis). However, as French and Steinberg pointed out, such results apply *in detail* only to the particular recording system, loudspeakers, listeners, and speech materials used. (Scharf, *op. cit.*)

If there is uncertainty regarding the role of critical bands in human speech perception, there must be even more regarding their relevance to the engineering of machine recognition. The issue has been obscured somewhat because of the running together of two things which are logically quite distinct, namely frequency-transformation on the one hand, and integration of spectral energies within bands on the other.

As regards clustering into critical bands (as opposed to any other bands), the statement I have highlighted in the last quotation from Scharf is worth focussing on. It is perhaps not coincidental also that the bands chosen in Holmes' 1980 vocoder (surely based upon practical experience) also increase more slowly with frequency (Owens 1993), and it was certainly my own experience that warping the frequency axis and clustering the energies to Zwicker's specifications led to sub-optimal recognition-performance, for the simple reason that the progression to wider bands approaching 2000 Hz was too steep and led to the obscuring of at least one and probably a number of phonetic distinctions. The most obviously affected distinction was that between [ii] and [uuf] (fronted [uu]). These two sounds have a more or less identical distribution of energy in the area of F1

(they are similar in HEIGHT), and while F2 in the case of [ii] is *usually* higher than in [uuf] (at 2000 to 2200 Hz, compared with 1700 to 1900 Hz for [uuf]), in particular instances, e.g. after [y], [uuf] may have F2 at or above 2000 Hz, and in these cases it is *only* the lower position of F3 in [uuf], at around 2300 Hz, that distinguishes the vowel from [ii] (which typically has F3 at about 2500 to 2800 Hz). With critical bandwidths of 320 Hz at centre-frequency 2150 Hz, and of 380 Hz at centre-frequency 2500 Hz, there is a danger of obscuring the distinction between these phones in some contexts.

On the other hand, there would appear to be an obvious truth to the statement that frequency-resolution needs to be finer at lower frequencies (in the regions crowded with the lower formants of vowels, for example) than at higher frequencies. Distinguishing 100 Hz bands throughout the higher frequency-region of an [s], for example, simply makes no sense for recognition-purposes. In general, it is hard to find fault with the idea that differences which a human ear is incapable of discerning are unlikely to be relevant in a communication system evolved by and for humans. It is not obvious, however, that differences in the degree of resolution required at different frequencies cannot be taken full account of by an appropriate frequency-transformation, without any emphasis being placed on the idea of bands as units of energy-integration. When deriving the cepstrum from a Fourier basis, however, frequency-transformation is effected *by means of* the clustering of individual FFT energies into bands, so that there is a question mark over whether the distinction (between frequency-transformation and energy-integration within bands) can actually have any meaning in practice.

The superiority of the Bark transformation to the linear scale would appear to be explicable on the basis of alignment (between DCT basis functions and the spectrum) alone. Given a frequency-range of 0 to 6 kHz, in the linear scale the first 4 coefficients do little more than provide discrimination between vowels and sonorants in one camp, and more or less everything else in the other, the difficult problem of vowel-discrimination not beginning to be seriously addressed until the 6th basis function is applied (the 5th basis function will have a broad peak at around 2500 Hz, which is not particularly significant for most vowels, and another at nearly 5000 Hz, which is virtually irrelevant); in subsequent basis functions up to the 10th, there is never more than one peak in the area of real interest

for vowels. With a Bark or similar transformation, the mid-point is closer to 1.5 than to 3 kHz (the region above 4 kHz being covered by a very small number of bands), with a great more attention paid in consequence to the region of the frequency-range of greatest significance for vowels.

We might perhaps get some purchase on the question of how significant energy-integration within bands could be by considering two schemes for deriving the cepstrum from the spectrum, call them A and B. Both schemes use a linear frequency-axis, covering (let us say) the range from 0 to 7500 Hz, but while in scheme A there is no banding — the spectrum is represented by 240 individual spectral coefficients — in scheme B the original 240 spectral coefficients are clustered into 48 equal bands each comprising 5 FFT coefficients. What sorts of difference may be expected between the cepstra produced under these two schemes? If we take any of the basis functions of the DCT (the reader may wish to refer back to figures 3.1 and 3.2) and imagine the cosine overlaid on the spectra for some vowel (say) under these two schemes, perhaps the most obvious point is that in scheme A the cosine will be 'sampled' at 240 points corresponding to the 240 spectral coefficients, while in scheme B it will be 'sampled' at only 48 points, each point corresponding to the centre-frequency of the frequencies covered by the particular quintet of spectral coefficients concerned. This in itself would cause fairly minor differences of outcome were it not that in scheme B the energies within each quintet are summed and the sum then logged before multiplication by the single cosine-value, whereas in scheme B we have, for the corresponding five spectral energies, five log operations before the multiplications by the cosine values; the products are subsequently summed as part of the summation of all the products for the particular basis function concerned. The log of a sum of 5 positive numbers yields a smaller value than the sum of the logs of those same numbers, so that absolute differences that may be quite substantial may appear in the cepstra produced under the two schemes. It is not obvious, though, why such differences should in themselves affect the discriminant abilities of the two representations so derived. Moreover, apart from the obvious differences in scale, the impression from the plots of the collected cepstra is one of considerable similarity overall (figures 3.3, 3.4, 3.7 and 3.8).

The importance of frequency-transformations can be illustrated by comparing the cepstra resulting from schemes A and B, and from two further frequency transformations, the Bark transformation ("scheme D") and one other to be detailed in a moment ("scheme C"). Below are plotted cepstra (first 10 coefficients) for two pairs of vowels — [a] and [e], and [ii] and [uuf] — under these four schemes. In scheme C, the frequency-scale is linear up to about 2400 Hz, with bands of 125 Hz up to that point; two 250 Hz bands take us to about the 3 kHz point, a further two 500 Hz bands to the 4 kHz point, and two 1000 Hz bands up to the 6 kHz mark. All the cepstra are from collections of automatically identified stable portions of the vowels (details of how the identification is effected will be given in the next chapter).

It can be seen that while schemes A and B appear to differ little from each other in discriminatory power (as far as the eye can tell from the plots), scheme C appears to offer greater highlighting of differences than either of them. The vowel cores for [a] and [e], for example, seem to be almost completely distinguishable in virtue of the third coefficient alone under scheme C, while under schemes A and B there appears to be massive overlap at every coefficient. The Bark transformation (scheme D) appears to provide better separation of [Ca] and [Ce] than either scheme A or scheme B, but is arguably a little less effective in doing so than is scheme C. In the case of [Cii] vs [Cuuf], scheme C again seems to provide the better separation as far as one can tell from the plots; for the second and fourth coefficients there is hardly any overlap under this scheme. These results are, of course, only suggestive, and it does not follow from the fact that certain vowels are more discriminable under one scheme than another that all vowels or all phones are so. Further, more rigorous tests of the efficacy of different banding and warping schemes are described in Chapter 6.

Cepstral coefficients do not have to be calculated from the Fourier spectrum, whether first frequency-warped or not, but Davis and Mermelstein found that Fourier-based mel-scale cepstral coefficients outperformed all of a number of leading contenders, in speaker-dependent recognition-experiments using template-matching via dynamic time-warping (Davis & Mermelstein 1980). The contenders included LPC-based mel cepstral coefficients, and cepstral coefficients derived directly from the Fourier magnitude spectrum (without prior clustering into critical

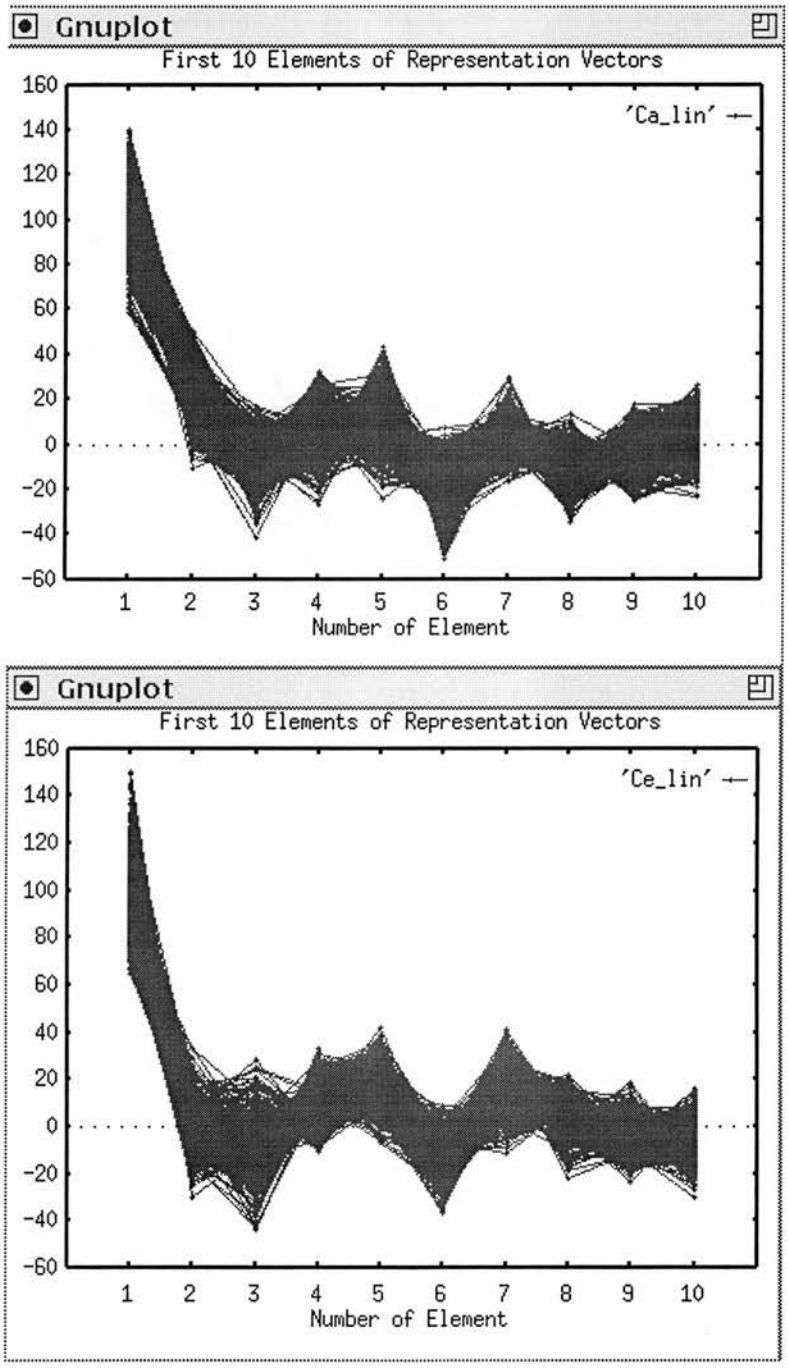



**Figure 3.3.** [a] vs [e] under scheme A

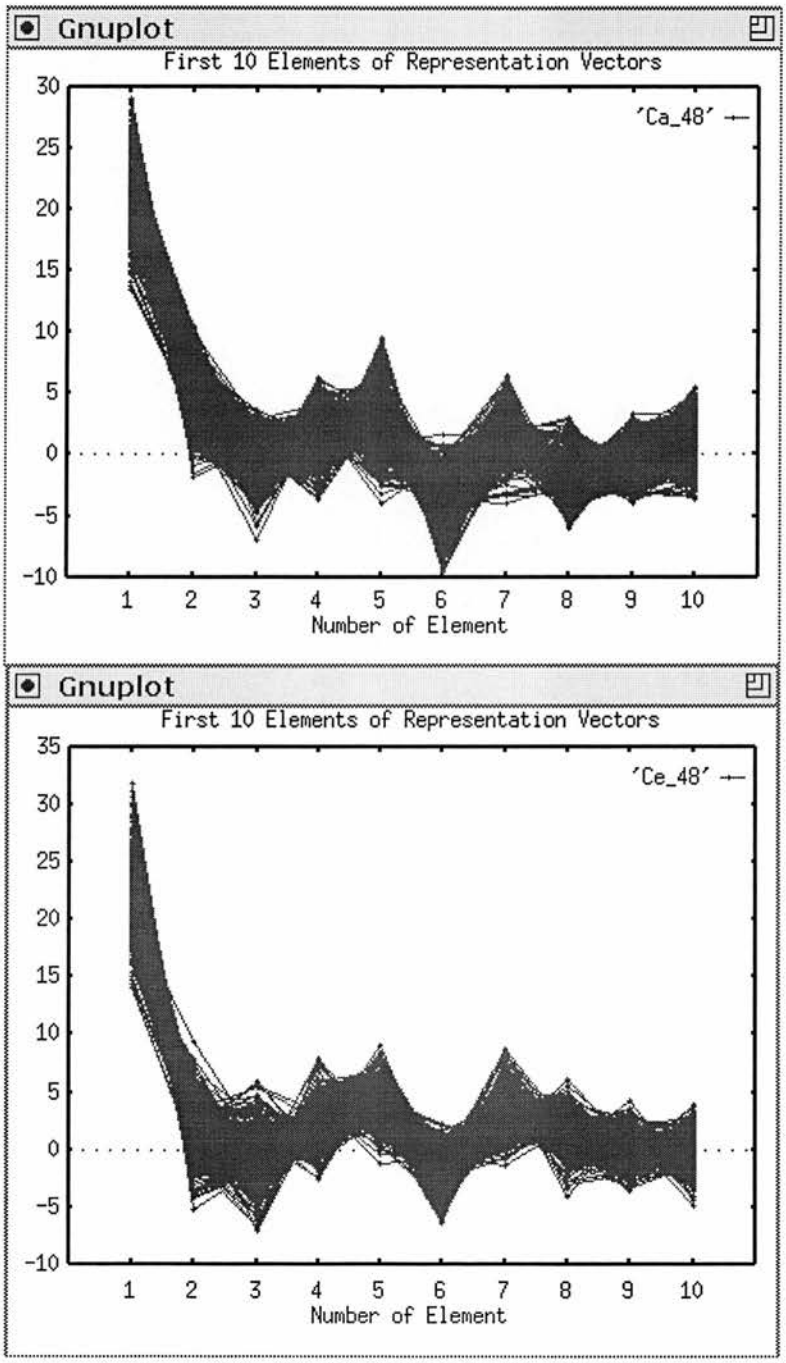



**Figure 3.4.** [a] vs [e] under scheme B

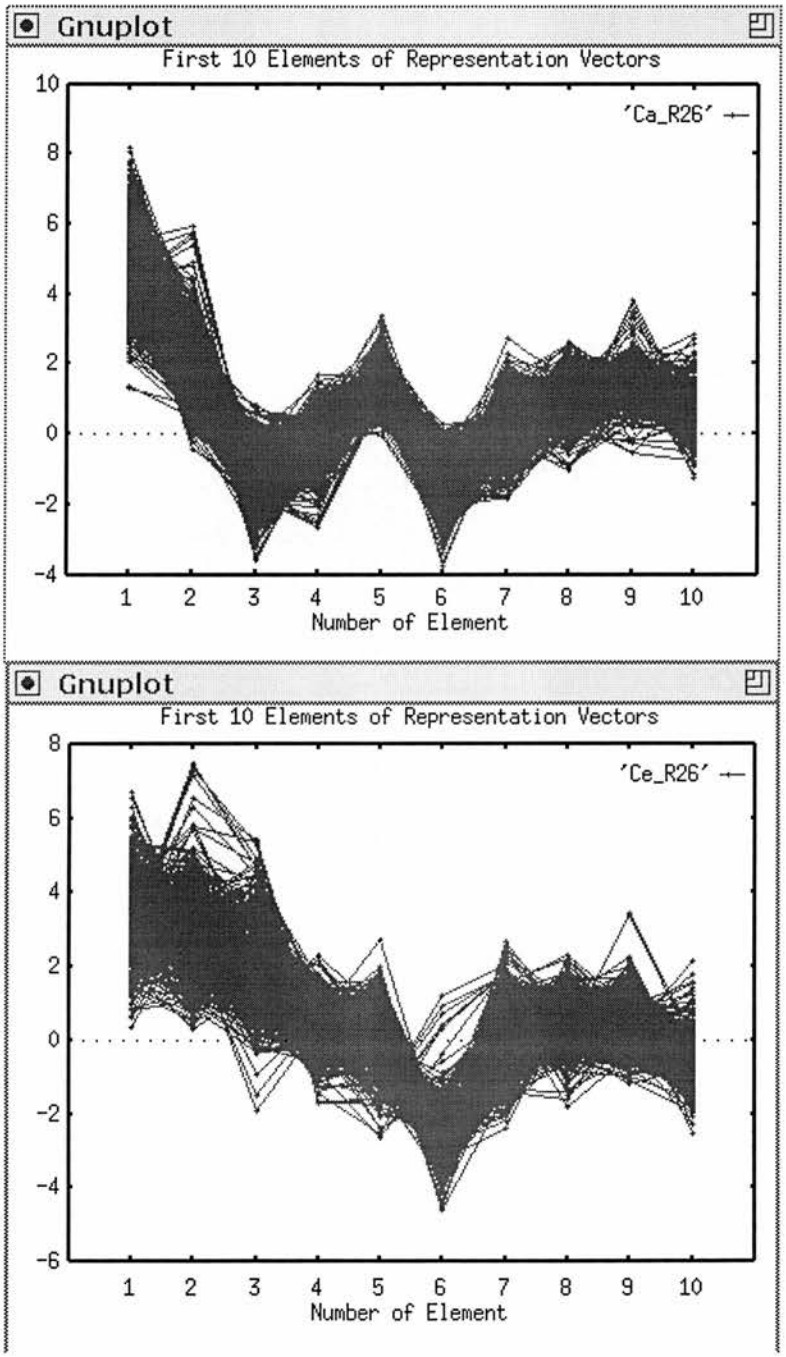



**Figure 3.5.** [a] vs [e] under scheme C

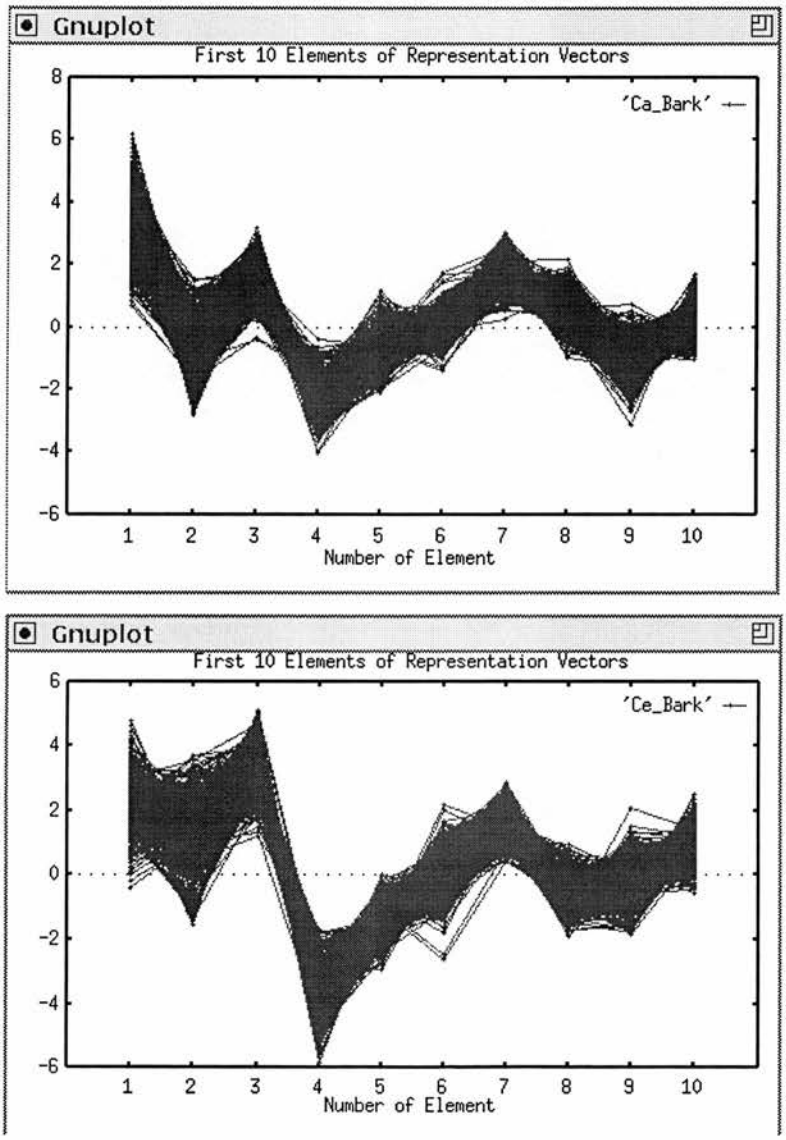



**Figure 3.6.** [a] vs [e] under scheme D

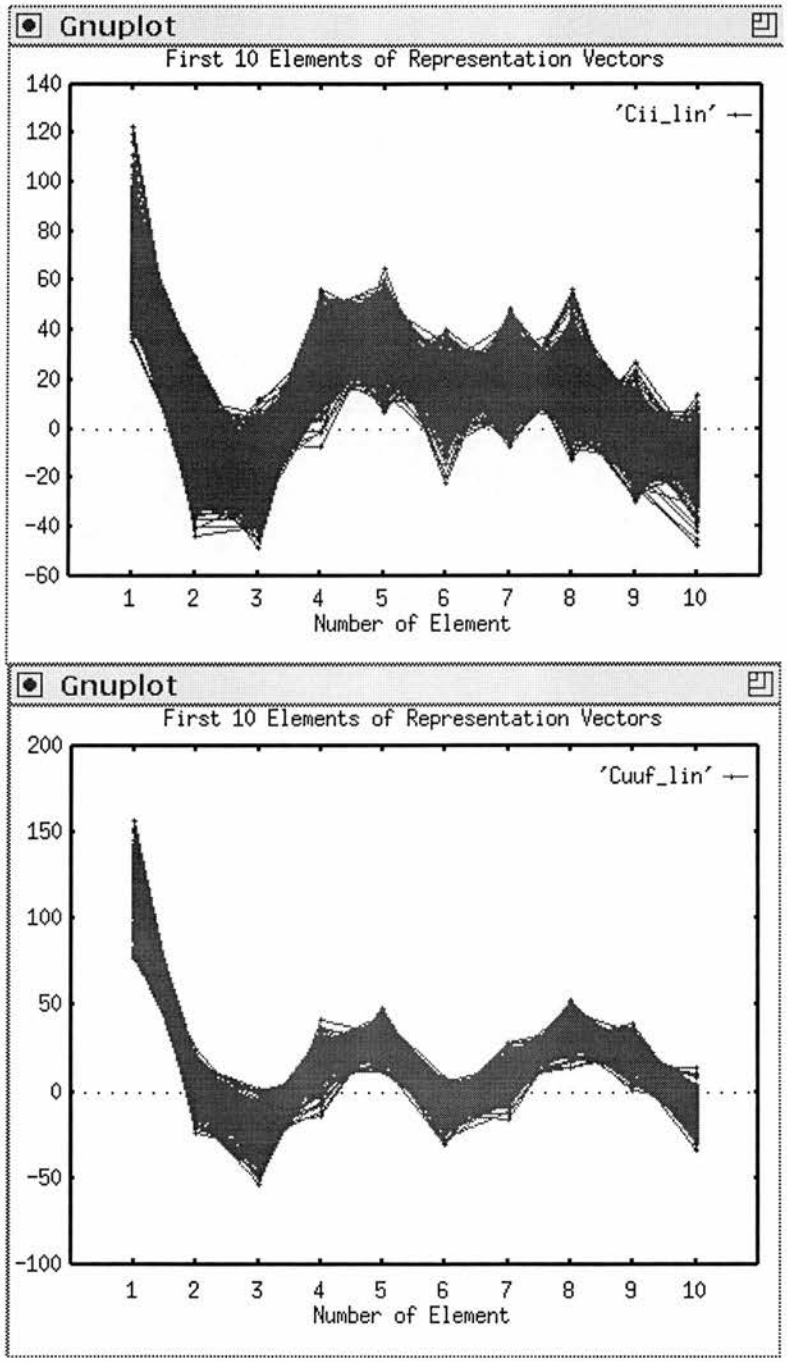



**Figure 3.7.** [ii] vs [uuf] under scheme A

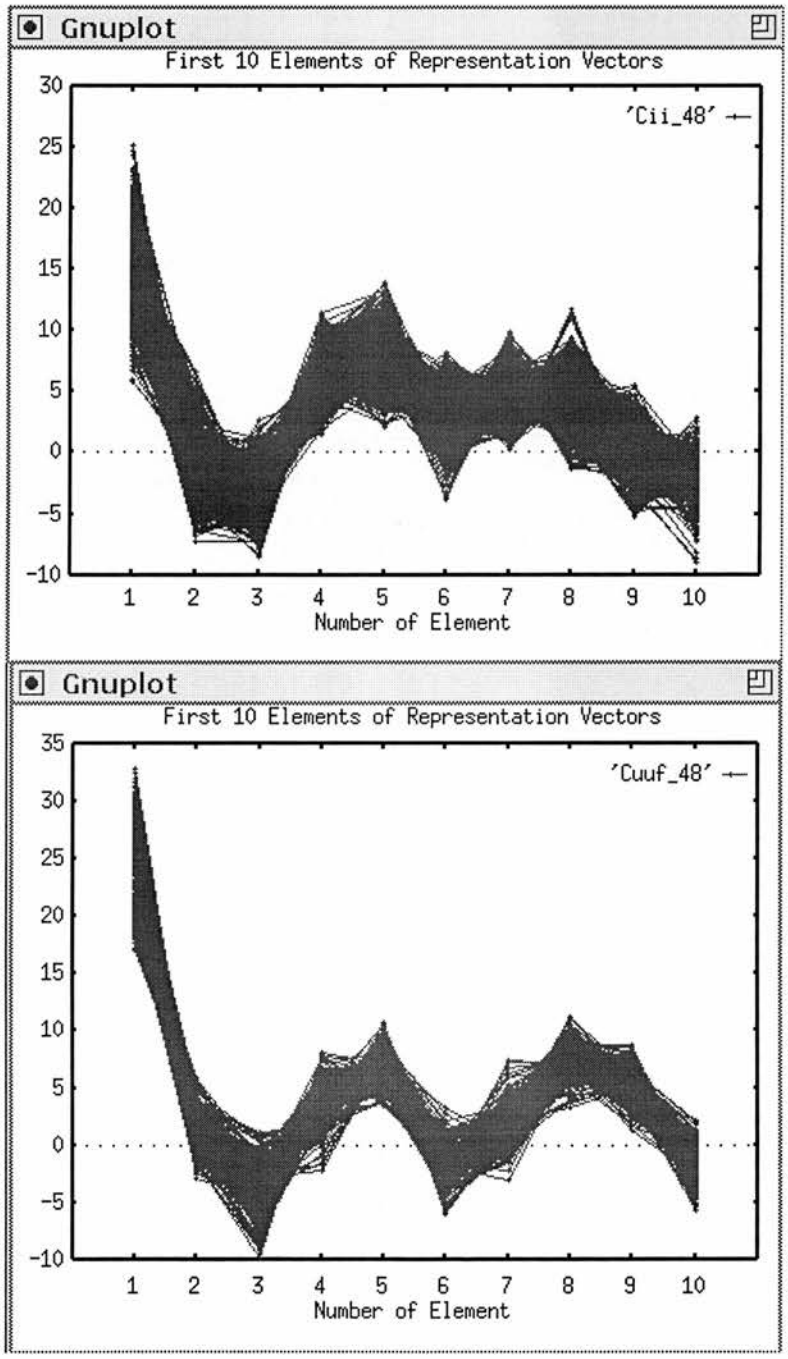



**Figure 3.8.** [ii] vs [uuf] under scheme B

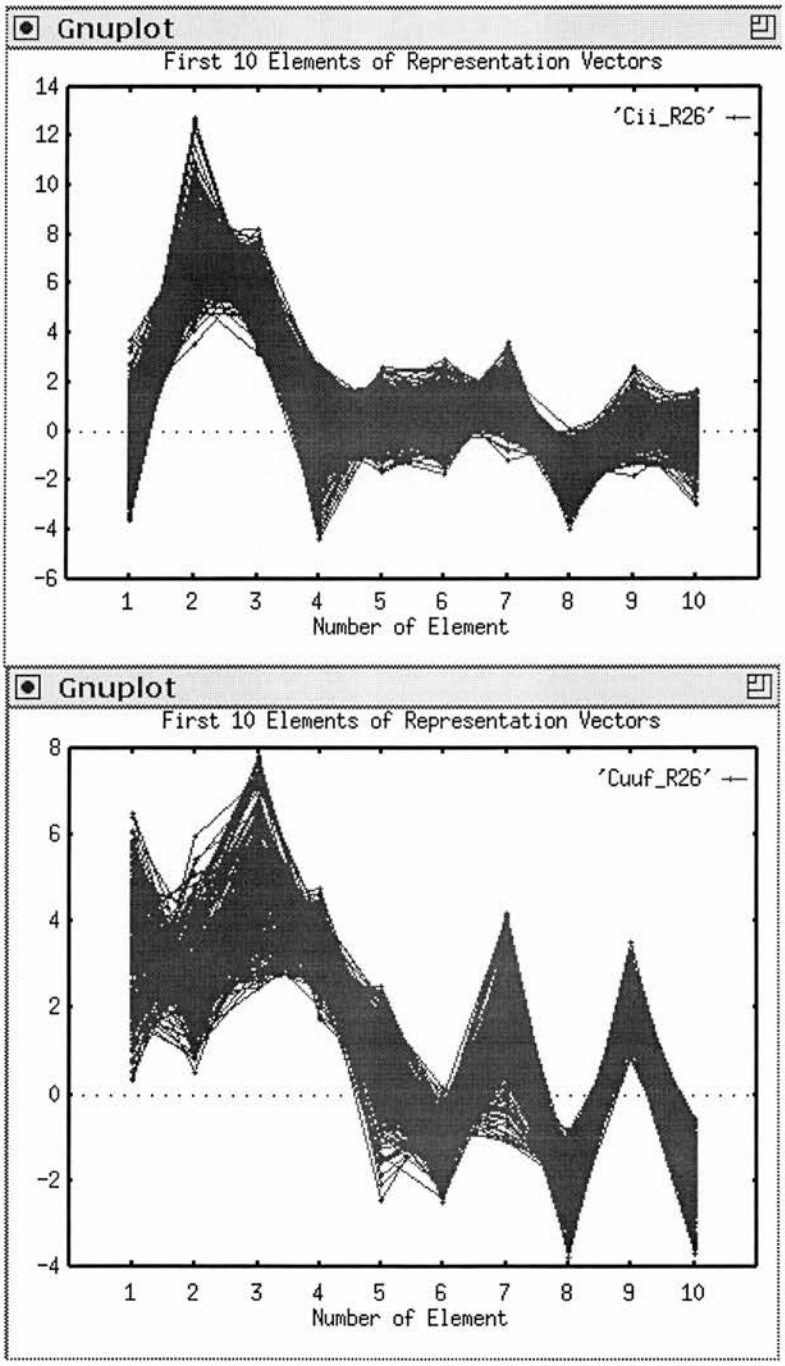



**Figure 3.9.** [ii] vs [uuf] under scheme C

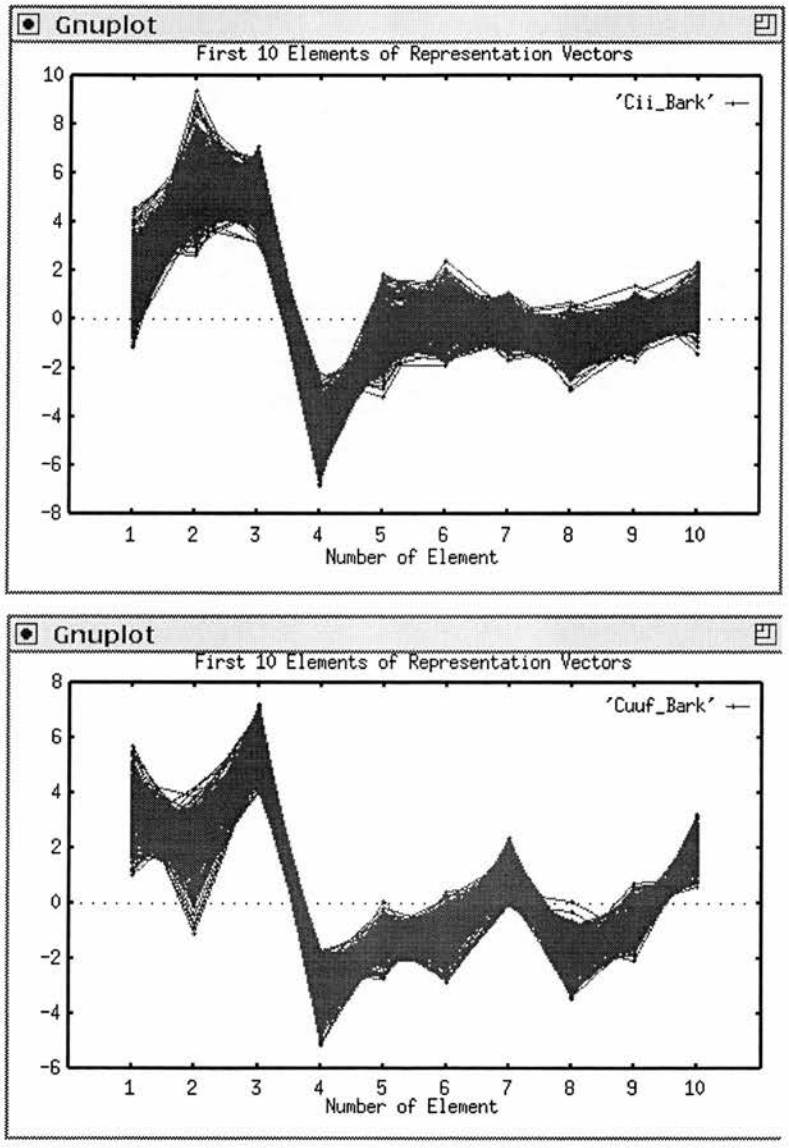



**Figure 3.10.** [ii] vs [uuf] under scheme D

bands). The inferior performance of the LPC mel cepstral coefficients may be attributable, as the authors suggest, to the poor modelling of nasalised speech using LPC, while the superiority of the mel-scale Fourier-based cepstral coefficients to those derived without frequency-warping and clustering is unsurprising with hindsight, for the reasons given earlier (though it should be noted that Davis and Mermelstein derived their linear scale cepstral coefficients from spectral coefficients covering just the range 0 to 4600 Hz).

In a series of preliminary experiments I compared the effectiveness of the best of the Davis and Mermelstein representations with a number of variants. These preliminary experiments involved statistical classification of eleven test-sentences (about 5000 vectors) using a Gaussian frame-classifier, and results were based on the average probability assigned to the correct class. In the first experiment, I deviated from their format just in respect of using a less steep increase in bandwidth, and found an improvement of 17% as a result (a 17% improvement, that is, in the average probability assigned to the correct phone). The Davis and Mermelstein representation involved the simulation of overlapping triangular filters, and in a second experiment using the same scaling as in the first experiment but with a banding-procedure corresponding to the use of rectangular, non-overlapping filters, some further improvement accrued, giving a 22% improvement over the results obtained following the Davis and Mermelstein specifications completely. Later experiments brought me to the scheme (scheme C) described above. Details of related experiments will be given in Chapter 6.

## 3.5 Window-Size and Stationarity – Border-Straddling ('TR') Classes

Davis and Mermelstein (*op. cit.*) found that an analysis window comprising some 25 ms of speech gave markedly better results than one comprising only half that figure. In early experiments based on classification of representation vectors from steady portions of vowels, I found that a 32 ms window yielded better results than a 16 ms window. It is a commonplace of the literature that a longer window is beneficial in the case of stable regions such as those associated with

vowel steady-states, but that a shorter window is required to capture sudden and short-lived changes in the signal, such as transients associated with stop-bursts and points of change between voiced and non-voiced phones (Schafer & Rabiner 1975). (The very fine temporal resolution of a typical broad-band spectrogram arises from use of an analysis window covering just 2 to 5 ms of speech). It was my original intention to develop a dual-window system to try to get the best of both worlds, but it turned out in tests of the single-window system with 32 ms window-duration that stop-bursts did not fare noticeably worse than any other kind of phone. It seemed easier to arrive at a specific remedy for gross violation of the stationarity requirement at many phone-boundaries than to work out a solution to the problem of how to integrate results from a dual windowing system, and this was therefore taken up as the next step. At that point, the development of the dual-window idea came to seem less of a priority and in the end the idea was never developed for lack of time. It is nevertheless true that relatively poor recognition-performance in respect of rapidly changing parts of the speech-signal may well be connected with the use of such a long analysis-window throughout.

It is no doubt fairly obvious that a long analysis-window is likely to yield strange results when the window happens to fall at a region of the speech waveform which comprises a sudden or dramatic change, at least where the change is not to or from silence. The spectra corresponding to these highly non-stationary regions may look very strange indeed, and including their representation-vectors in the training-data for any subphone — particularly for subphones with very meagre representation in the training-data — can have a significant distorting effect on the statistical parameters estimated for the class, while in recognition-mode the vectors are not particularly likely to score well with respect to their own class, particularly given the simple modelling scheme. On the other hand, simply excluding them from training-data altogether, while it removes the problem of distortion, leaves the problem of classifying them in recognition-mode unresolved. The policy adopted[3] was to model these border-regions with their own classes, so that for example there is a class for vectors derived from windows straddling [n],[e] boundaries, and similarly for [ii],[s] boundaries, and so on for all

[3]I am grateful to Fergus McInnes for the original suggestion

boundaries where the problem is capable of occurring. It is deemed unnecessary to employ such classes at boundaries which are strictly conventional, such as at borders between vowels, or between vowels and any of [r],[w],[y],[dl] (in either order), since in these cases the transition is gradual and the *precise* placement of label-boundaries cannot be taken too seriously, so that any window straddling one of these boundaries may fairly be allocated to one or other of the phones concerned; in these cases, if more than half of the windowed data lies on the left side of the label-boundary, the vector is assigned to the earlier phone (or subphone), otherwise to the later.

With rather limited amounts of training-data, it is not possible to model all the relevant border-regions with complete specificity, and generalised classes have to be used, modelling such things as [vowel],[nasal] and [[-VOICE] fricative],[vowel] borders. A full list of all these border-straddling classes (TR classes henceforth) is given in Appendix C. A few examples are provided here, following an outline of the criteria used for allocating vectors to such TR classes.

Letting 'TRxy' represent the class of frames straddling the boundary between phones of class x and phones of class y, for phone y with predecessor x and successor z, and considering frames whose centres are within the y segment:

1. assign the frame whose centre is closest to the centre of y to y;

2. assign any unassigned frame whose centre is within 8ms of a boundary to TRxy or TRyz according to which boundary it is closer to;

3. assign any still-unassigned frame to y.

This scheme [4] fails only when a phone is less than 6 ms long and has no frame-centres within it (there were no phones this short in any of the data worked with). Otherwise, the procedure ensures that there will be at least two and at most three TR-class assignments at any phone-boundary concerned.

An illustration of the use of TR classes is provided by those used to model boundaries between stop-closures and stop-bursts (these are labelled explicitly, as [tc],[tb] etc., as described in Chapter 2 (2.5.2)). Work by Stevens and Blumstein,

[4] I am indebted to Fergus McInnes for the algorithm, which proved better at coping with very short segments below 20 ms than the one of my own devising.

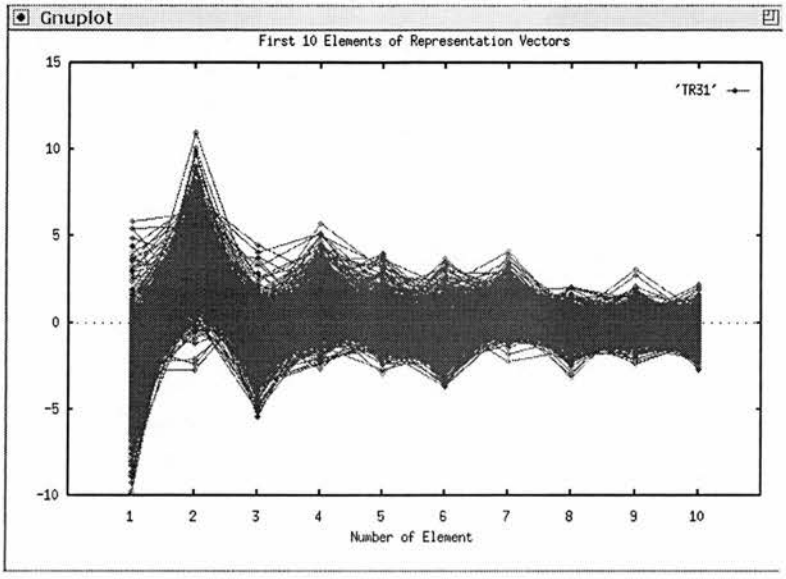



**Figure 3.11.** TR31: [tc]/[tb] borders

among others, suggests that the spectral properties of bursts associated with the moment of release of a stop are rather insensitive to context, and these TR classes are highly appropriate to such a state of affairs (Blumstein & Stevens 1979; Blumstein & Stevens 1980), though again the long anlysis window is no doubt unfavourable to accurate discrimination. Cepstra for the six main classes are plotted in figure 3.11 – figure 3.16.

## 3.6 Incorporating Trend Information

ASR systems normally supplement short-term information with information regarding the way the spectrum is developing over the slightly longer term. The first sort of information is encoded by so-called static features such as cepstral coefficients and power, the features encoding the latter sort then naturally being described as "dynamic" (also as *delta*) features. A variety of measures have been used for the dynamic features, including simple difference-measures (Lee 1990), and linear regression or rate-of-change coefficients (Lee *et al.* 1990a). For each cepstral coefficient used for the static representation, one calculates either

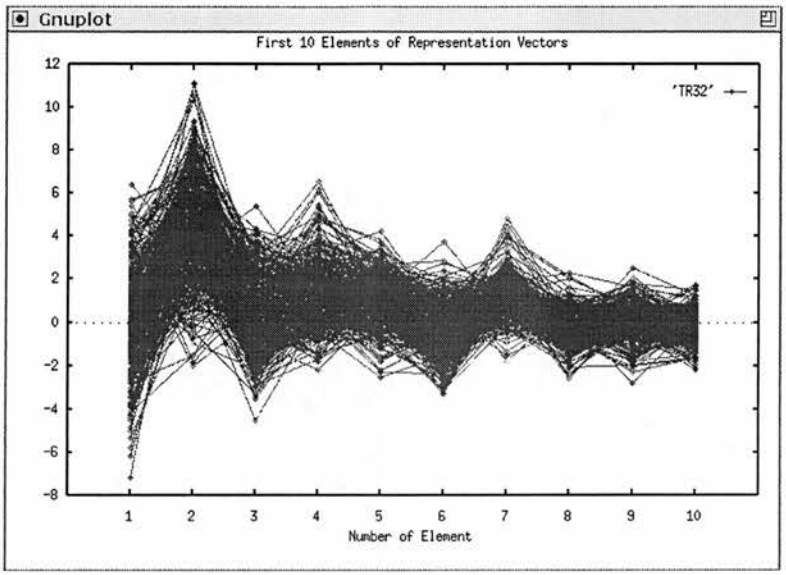



**Figure 3.12.** TR32: [dc]/[db] borders

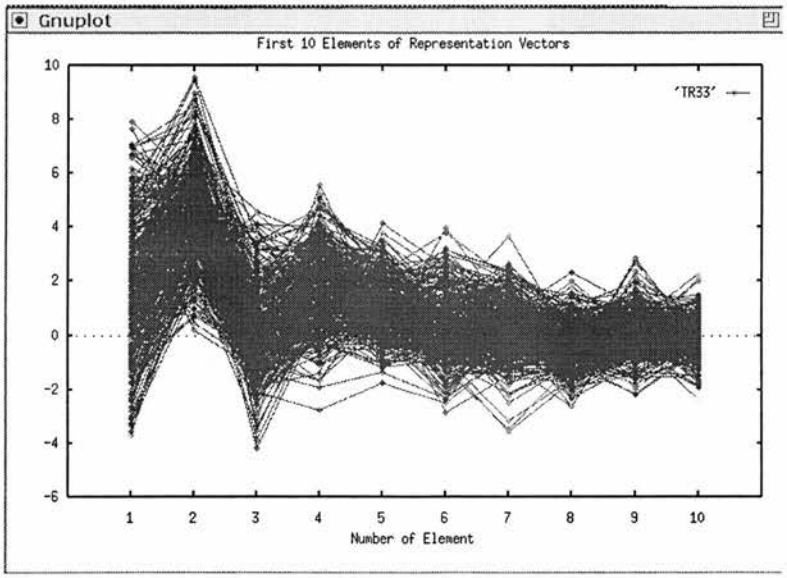



**Figure 3.13.** TR33: [bc]/[bb] borders

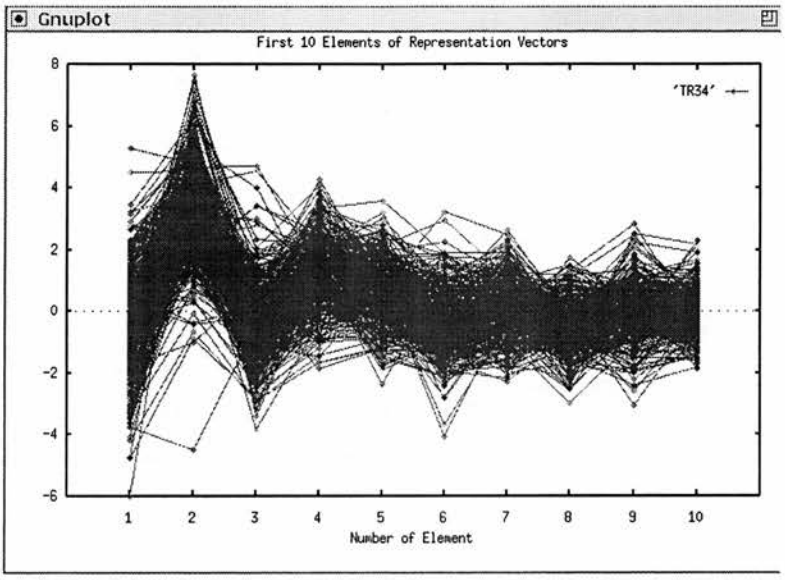



**Figure 3.14.** TR34: [pc]/[pb] and [pc]/[Pb] borders

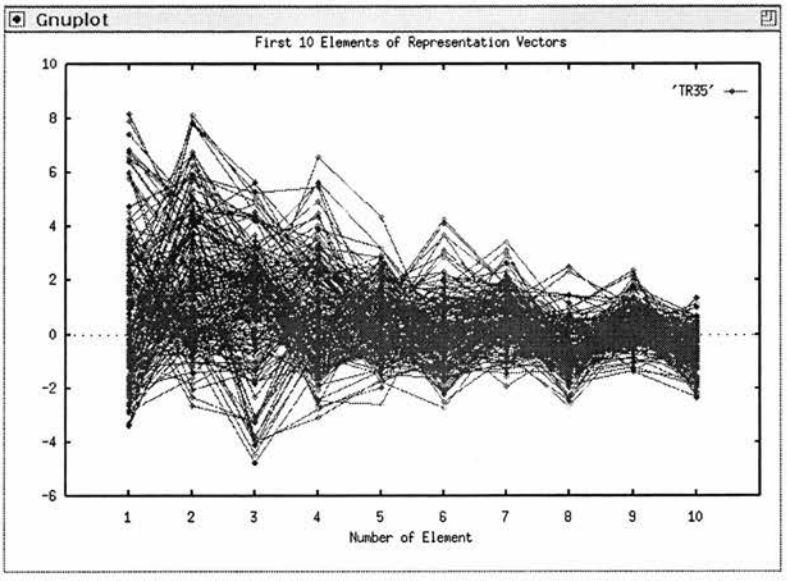



**Figure 3.15.** TR35: [gc]/[gb] borders

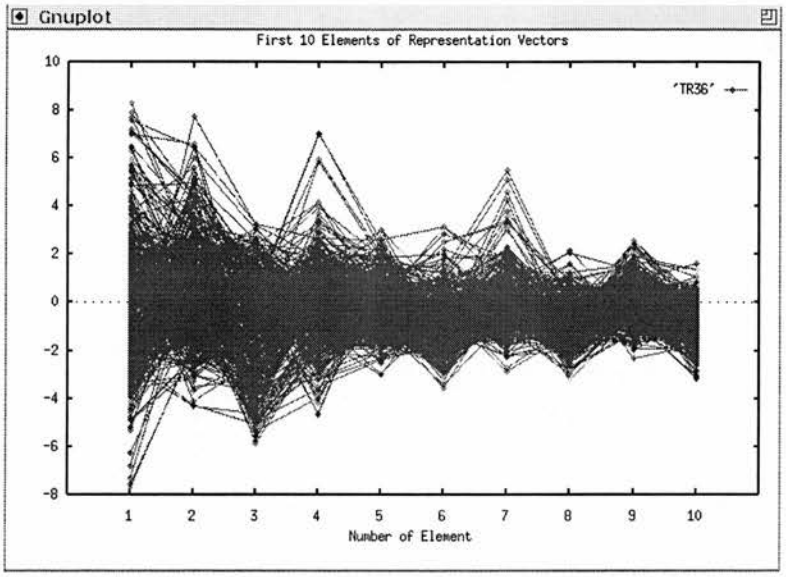



**Figure 3.16.** TR36: [kc]/[kb] and [kc][Kb]borders

the difference across some small number of frames centred on the current frame, or a measure of the trend in the values for that coefficient across that number of frames, thus producing a second vector which may be concatenated with the static representation vector.

Dynamic features used in this work were derived according to the formula given in the article by C.H.Lee *et al.*, calculated over 7 frames centred on the current frame:

$$\Delta cc_t(i) = \sum_{k=-3}^{3} k cc_{t-k}(i), \quad 1 \leq i \leq N, \tag{3.4}$$

where $cc_t(i)$ is the $i'th$ cepstral coefficient at frame $t$, and $N$ is the number of delta cepstral coefficients used. Normally I made use only of the first 4 delta coefficients; given that the cepstral coefficients are ordered with respect to their variance, those with greater variance coming first, and given the need to keep the dimensionality of the representation vector as low as possible, I initially worked with just the first 4 deltas on the assumption that the chances of still failing to disambiguate the class-affiliation of a vector with all the static and this number of dynamic coefficients must surely be quite small. Subsequent experiments

in which I used larger numbers of dynamic coefficients brought no significant improvement, given the need to make adjustments elsewhere (such as assuming statistical independence of all coefficients), so that the use of the subset of 4 deltas became standard for most of the work.[5]

It is probably rather obvious that dynamic information can help to disambiguate the phonetic identity associated with a particular frame of speech. Frames from the onset of a vowel V following some consonant C may, if attention is restricted to their static features, be difficult to distinguish from frames from the offset of the same V before the same C, while the inclusion of the dynamic features, which picks up on formant-trends over a number of frames, is likely to make it clear which of the two possible subphones the frames are actually associated with. But there are weaknesses in this way of representing trend-information. The measures are highly sensitive to speaking-rate and hence the values found in any instance may be misleading unless the statistical parameters are estimated from very large amounts of training-data, with data from across the full range of speaking-rates for each and every subphone we wish to model. With modest amounts of data it will often turn out that we have only 2 or 3 tokens for some particular (context-sensitive) subphone, and if the same subphone occurs in speech to be recognised spoken at a very different rate from those represented in the training-data, the dynamic coefficients will be positively unhelpful. Speaking-rate varies, of course, not only from occasion to occasion but even within the confines of a single sentence.

A second weakness arises not so much from the way the trend information is gained but from the use of the single gaussian to model the dynamic coefficients for a given class. Consider, for example, the dynamic coefficients for frames belonging to the onset of an [aa] following a [sh]: it is clear that with the long analysis-window for their calculation the coefficients will change quite radically

[5] In work carried out subsequent to the completion of work on this thesis, using the Cambridge HMM Toolkit HTK, I found that using only a subset of delta coefficients caused a very significant deterioration compared with results obtained using the full complement; this was with diagonal covariance matrices, and very large amounts of training data, so it is not safe to draw firm conclusions with respect to the present case; it may nevertheless be fair to suppose it likely that with much larger amounts of training data than that used in work for this thesis, and a full complement of delta coefficients, some significant improvement could be achieved over the results presented later in this thesis.

as the frame counter moves into the vowel and reaches a point where the last [sh]-frames fall out of view, while by the time the end of the onset is reached, half of the frames considered will typically belong to a steady state, so that the dynamic coefficients will be levelling off. It is hard to see how the various sets of dynamic coefficients can be seen as variations on a single theme, as the single gaussian model ideally requires.

## 3.7 Summary of the Representation Used

The representation of the speech-signal used throughout the development of the technique presented in this work was, then, as follows: ten cepstral coefficients were computed from the warped (log magnitude) Fourier spectrum using the warping described above as scheme C. The static coefficients were supplemented by the dynamic counterparts of coefficients 1 to 4, and by a log power term and its dynamic counterpart. The window-length for analysis was 32 ms, with the wondow advancing by 6 ms at a time, and the dynamic features were calculated over 7 frames centered on the current frame, using the formula given in the preceding section.

The search for a representation scheme started from the conclusions presented in the article by Davis and Mermelstein cited earlier, and the modification to what they found optimal was chosen on the basis of the somewhat better discrimination the modified scheme (scheme C) appeared to offer in preliminary experiments, as described in 3.4.3. The decision to use only the first four of the dynamic coefficients was prompted by the desire to keep dimensionality as low as possible given the limited amount of training-data, and because of the concentration of information in the lower-order cepstral coefficients. The fairly long window length was settled upon because of the advantages it offered in recognition of vowels in particular, and because classification of short phones or subphones such as stop-burst transients did not appear to be markedly affected by using a longer rather than a shorter window. It nevertheless seems likely that a dual windowing system, or something similar, would have offered advantages had there been time to develop it.

# Chapter 4

# Getting from Cepstra to Phones

## 4.1 Introduction

In this chapter I describe the process whereby statistical models are derived in FURIDA from sequences of pattern-vectors and associated phonetic labels, and show how these models are used to derive phonetic transcriptions from sequences of unlabelled pattern-vectors. I first describe these two processes in outline, and then focus on two topics relevant to both processes, namely statistical pattern classification and dynamic programming. Each of the two processes — training and recognition — is then described in further detail. I then compare the methods described here with HMM approaches, focussing in particular on the issue of transparency and on the question of phone-duration and its relevance to phonetic classification. Finally I consider the possibility — and desirability — of incorporating some form of duration-modelling in the system described in this work.

## 4.2 Training and Recognition Procedures in Outline

The goal of training is twofold: to learn the detailed subphonic structure of those phones in the training-data that are labelled at phone-level (and deemed to have

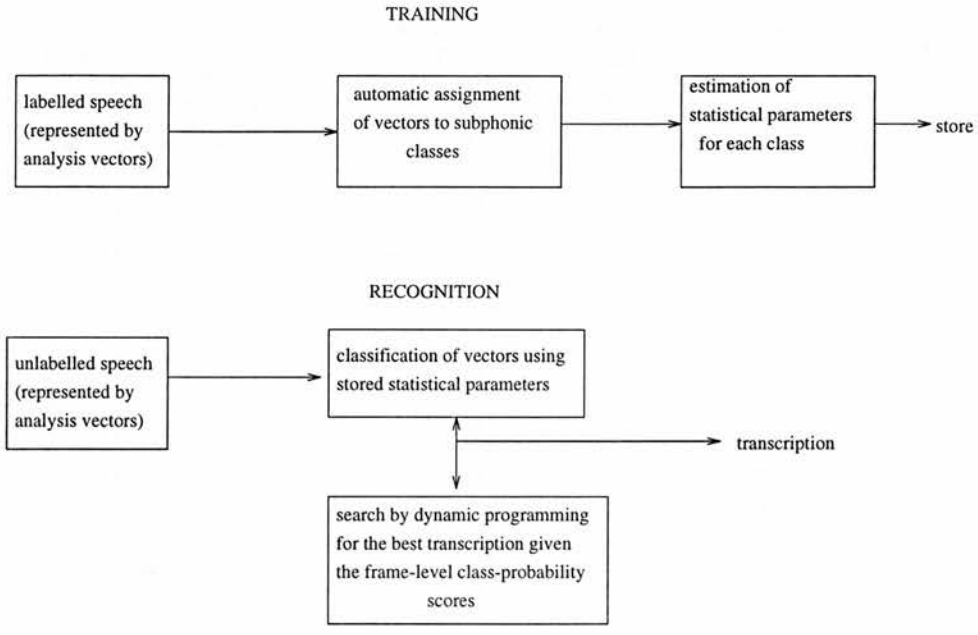



**Figure 4.1.** Training and Recognition Procedures in Outline

subphonic structure), and to derive statistical parameters for each (usually subphonic) class. In recognition, frame-level class-membership scores are calculated using the parameters estimated in training, and subphone-sequencing constraints are exploited to help trace the most probable overall transcription given the frame-level scores (figure 4.1).

It may be noted that in the basic system no attempt is made during training to learn subphone- or phone-durations explicitly (one rather minor qualification needs to be made to this, as will be explained in due course), nor to learn probabilities of transition between one subphone and any other within a phone.

In training, then, we start with speech labelled at phone-level for the very great part, but desire statistical parameters for subphonic classes in most instances, because of the sharper distributions this leads to and also because of the sequence-constraints it imposes on possible subphonic transcriptions. Accordingly, the training procedure begins with an initial attempt at dividing each targeted phone into subphones, and then seeks to improve on the first attempt

via an iterative process of closed-test reclassification[1]. In the first reclassification, the statistical parameters used are those derived after the initialisation, and in subsequent iterations the parameters used are those derived after the most recent iteration.

The initialisation for vowels, [dl] and [y] begins with the identification of the most stable quarter of the phone (excluding the first and final vector), which then becomes the basis of the phone's core or steady-state. Details of how this is done, and of the procedure whereby the rest of the phone is then allocated to onset, core or offset are deferred until sections 4.4 and 4.5. For most consonants, the initialisation is simply a 50:50 split between onset and offset.

Once all the training-data has been subjected to the initialisation procedure, the vectors for each subphonic class are collected together, and class-specific statistical parameters (mean vector and covariance matrix) are estimated from the class-specific samples. The process of reclassification then begins. In each iteration, each training utterance is revisited, and each phone within it that is deemed to have subphonic structure is resegmented internally in the light of the current set of relevant statistical parameters. The resegmentation of each phone proceeds as follows: class-membership scores are computed for each of the possible subphonic classes (usually three for vowels, [dl] and [y], and two for other consonants) at each vector of the phone, and the best-scoring path through the resulting matrix of class/frame scores (subject to certain sequencing-constraints) is then taken as the most probable subphonic resegmentation. By processing all the utterances in the training-set in this way, a new set of class-specific samples is obtained, from which new statistical parameters can be estimated at the end of the iteration, for as long as it seems desirable to continue the process. Generally it was found that very little difference was made by continuing beyond three or four iterations.

In recognition, the final set of statistical parameters is used to derive frame/class-membership scores in a similar way to that followed in training, except of course that in recognition, class-scores are potentially required for all of the classes at

[1]In a closed test, classification is performed on data that itself formed part of the samples used to estimate the statistical parameters; in an open test, by contrast, classification is performed on data that was not included in such samples.

every frame. The tracing of the most probable transcription for the utterance as a whole is again similar to the related process in training, though with complications arising from the presence of TR classes (section 3.5), and from the greatly more intricate set of sequence-constraints. Further details will be given in section 4.5.

## 4.3 Statistical Pattern Classification

Frame-level class-membership scores play a central part in both the training and recognition procedures. In this section I explain the fragment of statistical decision theory that is used in the calculation of these scores. (The account is chiefly based on chapters 1 to 3 of the classic work by Duda and Hart (Duda & Hart 1973) and chapters 1 to 3 of Schalkoff's book (Schalkoff 1992) (which largely precis Duda and Hart).) (A wonderfully lucid exposition of the material is now available in the early chapters of (Bishop 1995).)

The idea of characterising an object in terms of a set of measurements is a familiar one to any student of Phonetics, where vowels for example are often represented by scatter plots of F1 against F2 (or some other relationship between formant-values). Given such a representation, it becomes a matter of interest to describe the way the measurements for the two features are distributed, both absolutely and with respect to each other: how much variation is there in F1, and in F2, and around what mean values? What kind of correlation, if any, is there between the two sets of measurements? The answers to these questions may be presented in the form of a mean vector (giving the mean value of F1 and the mean value of F2) and a covariance matrix (giving in the off-diagonal cells the unnormalised correlation between the features, and in the diagonal cells their variances), and for a normally distributed population these two parameters (the mean vector $\mu$ and the covariance matrix $\Sigma$) specify the distribution fully. [2] An example set of parameters for the vowel [ii] for some hypothetical speaker might be as follows:

[2] A vector is usually to be thought of as a column vector as here, but may be represented in text in the form $(x_1, x_2)$. In formulae, the notation $\mathbf{x}^t$ indicates the transpose of the column vector $\mathbf{x}$; thus if $\mathbf{x}$ = the column vector (480, 1740), $\mathbf{x}^t$ = the row vector (480, 1740).

$$\mu = \begin{bmatrix} 339 \\ 2041 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 14453 & -3149 \\ -3149 & 30251 \end{bmatrix}.$$

Note that in the covariance matrix the upper triangle comprising the diagonal and everything above it mirrors the lower triangle comprising the diagonal and everything below it, the covariance between F1 and F2 (or in general between $x_1$ and $x_2$) being precisely the same thing as the covariance between F2 and F1 (or $x_2$ and $x_1$)! This remains true, of course, regardless of how many features are represented in the matrix.

It is of course essential to distinguish between values for a sample of measured vowels (or a summary of a set of such values by means of a sample mean vector and sample covariance matrix) and the true or population values and parameters. It goes without saying that what we would like for present purposes is accurate estimates of the population parameters, since we are intending to score vectors for membership of different classes using these estimates. What we in fact have, of course, are only samples, and we have to use the samples to estimate the population parameters. One method for calculating these estimates, the method to be used here, is the method of Maximum Likelihood (ML), which takes those estimates as true which give the greatest probability to the samples. (It is a crucial if obvious point that if a sample is unrepresentative, classification accuracy using the ML estimates based upon it will reflect this.) For data assumed to be normally distributed, the ML estimates for $\mu$ and $\Sigma$ are

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \tag{4.1}$$

and

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t. \tag{4.2}$$

Thus the best ML estimate for the population mean vector is the mean of the sample vectors, and that for the population covariance matrix is the average of the n matrices $(x_k - \hat{\mu}) \times (x_k - \hat{\mu})^t$. [3] The ML estimate for the covariance matrix

[3] To multiply $(x_1, x_2)$ by $(x_1, x_2)^t$, multiply $x_1$ by $x_1$ and put the result in $cell_{1,1}$; then

is biased, but the bias can be corrected by substituting $\frac{1}{n-1}$ for $\frac{1}{n}$ in (4.2) [4].

It is helpful to return to our F1,F2 scatter plot at this point and to try to imagine a plot which gives values for an entire population (say, all the tokens of /ii/ ever produced by some individual who was recorded every time he ever spoke). We should expect to find certain areas of the F1,F2 space to be thoroughly saturated with data-points, and others to be less sparsely populated, with the general form of a circle or ellipse emerging if the data is normal. The use of the term *density* to characterise the lie of the measurement-values needs no explanation. A probability density function (*pdf*) defines, for any set of measurements $\mathbf{x}$ (e.g. a pair of F1,F2 values) the probability of $\mathbf{x}$ having come from a particular population of measurements (and therefore from the class represented by those measurements). For normally distributed data, we have

$$p(\mathbf{x}) = \frac{1}{2\pi^{d/2} \times \mid \Sigma \mid^{0.5}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1}(\mathbf{x}-\mu)\right] \tag{4.3}$$

where $p(.)$ is the probability density function, $d$ is the dimensionality of the representation vector $\mathbf{x}$, $\mid \Sigma \mid$ is the determinant of the covariance matrix, and $\Sigma^{-1}$ is the inverse of the covariance matrix. [5]

We make the class-dependence of $\mathrm{p}(\mathbf{x})$ in (4.3) explicit by writing $p(\mathbf{x} \mid c_i)$ (the probability of $\mathbf{x}$ given that the class is $c_i$) and subscripting the other class-specific terms accordingly. Bayes' Theorem then tells us how to derive the probability of class $c_i$ given the representation vector $\mathbf{x}$:

$$P(c_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid c_i)P(c_i)}{p(\mathbf{x})}, \tag{4.4}$$

---

multiply $x_1$ by $x_2$ and put the result in $cell_{1,2}$; then multiply $x_2$ by $x_1$ and put the result in $cell_{2,1}$, and finally multiply $x_2$ by $x_2$ and put the result in $cell_{2,2}$. This should make clear that the diagonal of the covariance matrix will hold the familiar squared deviations from the mean.

[4] An estimator is said to be biased if the mean of its sampling distribution is not equal to the unknown parameter ($\Sigma$ in this case)

[5] The determinant of a $2 \times 2$ covariance matrix equals the product of the variances minus the product of the covariances ($\det(A) = a_{1,1}a_{2,2} - a_{2,1}a_{1,2}$). In general (regardless of dimensionality) the determinant of the covariance matrix is directly related to the scatter of the class data (Duda & Hart 1973). The product of a covariance matrix $A$ and its inverse $A^{-1}$ is the identity matrix $I$, which has 1's in the main diagonal and 0's everywhere else. A matrix has an inverse only if it has a non-zero determinant, and a matrix with a zero determinant is said to be singular. A matrix is said to be *positive definite* when it has a determinant greater than zero.

where the unconditional density $p(\mathbf{x})$ is given by

$$p(\mathbf{x}) = \sum_{j=1}^{n} p(\mathbf{x} \mid c_j) P(c_j). \tag{4.5}$$

The left hand side of equation 4.4 represents the *posterior probability* of the class $c_i$ given the vector $\mathbf{x}$, while $p(\mathbf{x} \mid \mathbf{c_i})$ is referred to as the *acoustic likelihood* of $\mathbf{x}$ given the class $c_i$. $P(c_i)$ is the a priori probability of class $c_i$ occurring. The a priori class probabilities may be estimated by calculating the relative frequency of occurrence of each class in the training-data, though in the present work, the a priori probabilities are treated as equal and so effectively ignored; the original motivation for this decision was that since the training-data was artificially designed to give each possible phonetic context for any phoneme as much representation in the data as possible, any a priori probabilities estimated from it would be unlikely to reflect any natural property of English phonetics or phonology. (In practice, once subphones have been derived, and generalisation of subphones has been effected to cope with data-shortages (see Chapter 5), there are sometimes very dramatic inequalities between the likelihood of occurrence of different classes; at the time of writing, however, it is still the case that no differentiation has been made between a priori probabilities; experience elsewhere suggests that inclusion of the class priors would bring about a significant improvement to the performance reported later in this work.)

The denominator in (4.4) is a normalising term designed to ensure that the posterior probabilities $P(c_i \mid \mathbf{x})$ will sum to 1 across all the classes. Since in the present context we are concerned rather with the ranking of scores than their absolute values, this term too can be ignored, so that we rely solely on the class-conditional scores $p(\mathbf{x} \mid c_i)$.

Bishop provides a useful formula for remembering the form of Bayes' Theorem (Bishop 1995):

$$posterior = \frac{likelihood \times prior}{normalisation factor}. \tag{4.6}$$

In statistical pattern classification, it is frequently the case that a classification follows immediately upon the calculation of class-membership scores, and the scores are thus naturally spoken of as *discriminant scores*, and the functions

that yield the scores as *discriminant functions.* Since ranking of the scores is the important factor, any function which preserves the ranking will serve the classificatory purpose equally well, and it is convenient to take as a discriminant function

$$g_i(\mathbf{x}) = \log\{p(\mathbf{x} \mid c_i)\}, \tag{4.7}$$

(the log function meeting this requirement), with (4.3) now giving way to

$$g_i(\mathbf{x}) = -0.5(\mathbf{x} - \mu_i)^t \Sigma_i^{-1}(\mathbf{x} - \mu_i) - d/2(\log(2\pi)) - 1/2 \log \mid \Sigma_i \mid . \tag{4.8}$$

The class-independent term $(d/2)\log(2\pi)$ can in turn be dropped, leaving us with

$$g_i(\mathbf{x}) = -0.5(\mathbf{x} - \mu_i)^t \Sigma_i^{-1}(\mathbf{x} - \mu_i) - 1/2 \log \mid \Sigma_i \mid . \tag{4.9}$$

It is convenient for purposes of computation to take the negative of $g_i(\mathbf{x})$, so that the best ranking class will be that with the lowest discriminant score.

In the present application, classification does not depend solely on the score at a single frame, but involves taking into account the scores over several frames. In the next section I turn to consider the methods used for doing this.

## 4.4 Dynamic Programming

Small-scale dynamic programming (DP) is used in the course of training, to help determine the most probable subphonic segmentation of individual phones, with no more than three classes and usually somewhere between 5 and 30 frames involved in any one case. In recognition, DP is used on a much larger scale to trace the most probable phonetic transcription of an entire utterance, with typically (in the work done for this thesis) about 750 phonetic classes and between 500 and 1000 frames involved in the search. The application of DP is essentially the same in both cases, in spite of the differences in scale. (I am indebted here to (Smith 1991).)

To find the most probable subphonic segmentation of a phone, given the class-membership scores for each subphone at each of the phone's frames, we may represent the data in the form of a matrix of at most three rows (for onset,

core and offset) and N columns (where N of course equals the number of frames), and cast the problem in the form of a simple path-finding exercise: what is the best-scoring path through the matrix that conforms with our ideas of what is reasonable from a phonetic point of view? (The question of what the precise constraints should be is fairly simple for the training context, and will be revisited a little further below; for now, it may simply be stated as an example of what is meant that we will at the very least require a path which never reverts to onset once it has passed into core, and never reverts to core or onset once it has passed to offset.) A problem formulated in this way may be efficiently solved using dynamic programming.

Once the idea of DP has been grasped it is actually very simple, but elaborate accounts of it can be confusing to the newcomer (if only because, after so much intricacy in the example used to teach it, the student thinks the idea must really be something quite profound or abstruse!). The essence of the matter is therefore stated here first in the simplest terms: in order to find the best path in the problem just described, we do our calculations in a particular order which enables us to re-use the results of earlier calculations, and to avoid repeating calculations already done. The contribution of Bellman, who first developed the technique, was to provide a key for the working out of what that order should be, in the form of his Principle of Optimality — the best policy is made up of the best sub-policies. Given that the best path from frame 1 to frame $N$ is the path we seek, we may focus on the notion of *sub-paths within the best path* which run from the first frame to some later frame. Thus if a path from 1 to $N$ defined by the class sequence

*AAAABBBCCCCC*

is the best path overall, then

*A*,

*AA*,

*AAA*,

*AAAA*,

*AAAAB*,

*AAAABB*,

*AAAABBB*,

*AAAABBBC*,

*AAAABBBCC*,

*AAAABBBCCC*,

and

*AAAABBBCCCC*

are all sub-paths of the best path in the sense intended. One essential if perhaps obvious point to grasp is then that any sub-path of this kind is itself a best path *to the point at which it terminates*; (there may be better-scoring paths to other class terminating-points, but that is immaterial to the present point). It turns out that the most efficient way to find the best overall path is by finding the best sub-path to *each* class at each frame, and this can be done efficiently by simply building on (extending) the sub-paths to each class that were found to be optimal at the preceding stage. When we reach the final frame, the best path of all the best paths to any class is the path we seek.

Letting $j$ index the stages (frames) in the search, and $i$ index the classes involved at each stage, one way to conduct the search is as follows: at $j = 1$ the possible sub-paths at each $i$ may be set to the class-scores of the $i, j = 1$ themselves. At $j = 2$, we take the best-scoring class $k, j = 1$ accessible from each $i, j = 2$, and set the sub-path score for paths to each $i, j = 2$ to the sum of the scores of the two classes ($i, j = 2$ and its best $k, j = 1$) that constitute the path. Generalising now beyond just the first two frames: for *each* $i, j$ we identify a class at the preceding column that constitutes the best predecessor (in the sense of terminating the best sub-path to itself), and we record both the index of that class and the sub-path-score to $i, j$ through that predecessor. The crucial point to note is that once we have done these calculations at $j = 1$, (and equally at

$j = 2$, $j = 3$, $j = 4$, ...), the results remain valid throughout the rest of the search: once we have completed work at stage $j$ we will always know the best predecessors of any $i, j$ (whatever $i$ or $j$ may be). When we have extended all the optimal sub-paths to stage $N$, we take the best scoring one as our solution, using the recorded best predecessors of each $i, j$ it comprises to recover the desired class-sequence (the most likely subphonic segmentation in the present case).

We define a 'predecessor' function (abbreviated below to 'pred') that states for each class the set of classes that may legally precede it in a path.[6] For vowels, [y] and [dl] the predecessor function in training is:

$$\begin{aligned} pred(onset) &= \{onset\} \\ pred(core) &= \{onset, core\} \\ pred(offset) &= \{onset, core, offset\}. \end{aligned}$$

For most consonants the function is:

$$\begin{aligned} pred(onset) &= \{onset\} \\ pred(offset) &= \{onset, offset\}. \end{aligned}$$

(In recognition, predecessor functions are defined across as well as within phone-boundaries, of course, and each class has a unique predecessor function.) Then we may express the recurrence relation at the heart of the DP search procedure as

$$Pscore*(i,j) = \min_{k \epsilon pred(i)} Pscore*(k, j-1) + score(i,j), 1 < j <= N, \quad (4.10)$$

(where $Pscore*(i,j)$ designates the score of the best path to $i, j$) with

$$Pscore*(i, j = 1) = score(i, j = 1) \quad \forall i. \quad (4.11)$$

[6] Clearly, since we are dealing with classification-scores for frame-based acoustic vectors, talk of classes being able to precede each other must be undertood as referring to classification-sequences: we are stating, by means of the predecessor functions, which sequences of classifications are legal and which are not.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| onset | 5.4 | 3.1 | 4.7 | 7.2 | 7.1 |
| core | 3.2 | 5.9 | 2.9 | 1.8 | 3.4 |
| offset | 4.8 | 7.2 | 5.4 | 4.2 | 3.1 |

**Table 4.1.** Matrix of Class/Frame Scores for Five-Frame Vowel

I conclude this section with a simple worked example, illustrating the use of dynamic programming to find the best subphonic segmentation of a vowel given frame-level scores for the subphone-classes involved. The reader is referred to table 4.1, which shows the scores for onset, core and offset subphones at each frame of a five frame vowel; the reader is asked to assume that these are negative log probability scores, so that the object of the DP search is to find the least costly path from the first frame of the vowel to the last. As we work our way through the table, frame by frame, we record in table 4.2 the scores of best sub-paths to each class at each frame, and the best predecessor for each class at each frame. Note that we outlaw starting in offset or ending in onset.

At frame 1, the best paths from onset and from core have scores equal to the frame-level scores of these classes. At frame 2, we begin our calculations with onset. Its only legal predecessor at frame 1 is onset, so we immediately extend a sub-path from onset at frame 1 to onset at frame 2. This sub-path has a score of $5.4 + 3.1 = 8.5$, which we record in the appropriate cell of table 4.2, together with the identity of the best predecessor (in brackets) for that cell. We then go on to core (cell 2,2 of table 4.1); its legal predecesors are onset and core and the best sub-path to extend to core at frame 2 is that from core at frame 1; this path when extended has a score of $3.2 + 5.9 = 9.1$. We record this sub-path score and also the identity of the best predecessor (core). Offset at this stage has only core and onset as its legal predecessors, and inspection of the scores in table 4.1 reveals that the best sub-path to extend to it is that which begins in core; this sub-path has a score when extended of $3.2 + 7.2 = 10.4$. We record the sub-path score and the identity of the best predecessor (core), and are then finished at frame 2.

We then go on to frame 3, and repeat the process of finding the best-scoring sub-paths to extend to each of the classes at that frame, using the results already

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| onset | 5.4 | 8.5(onset) | 13.2(onset) | 20.4(onset) | |
| core | 3.2 | 9.1(core) | 11.4(onset) | 13.2(core) | 16.6(core) |
| offset | | 10.4(core) | 14.5(core) | 15.7(core) | 16.3(core) |

**Table 4.2.** Accumulated Best Sub-Path Scores, and Best Predecessors in Brackets

established for best sub-paths to each class at frame 2. The process is repeated until all frames have been covered.

Inspection of the final column of table 4.2 reveals that the path ending in offset is the best-scoring (lowest-scoring) path in this case, so we begin our traceback from there: the best predecessor we go to is core at frame 4, whose best predecessor at frame 3 is recorded as core, whose best predecessor at frame 2 is recorded as onset, whose best predecessor at frame 1 is recorded as onset. Hence the best subphonic segmentation in the light of the frame-level scores is

```
ONSET ONSET CORE CORE OFFSET.
```

## 4.5 Further Details of the Training Procedure

### 4.5.1 The Initialisation

The following groups of phones are treated differently from one another, for reasons which relate to the phonetic characteristics of such phones discussed in detail in Chapter 2.

- monophthongal vowels, elements of two-phase diphthongs ([au], [ou], [i@], [e@], [eH@], and [u@]), [dl] and [y] (this group will be referred to as group 1 for the rest of this section);
- first-phase elements of three-phase diphthongs ([ai], [oi] and [ei]) preceding cognate diphthongal glides ('group 2');
- phase-three elements of three-phase diphthongs ('group 3');

- glide elements of three-phase diphthongs ('group 4');
- other consonants ('group 5').

The differences between the initialisation-procedures for the first three groups arise entirely from their conception or definition as phonetic entities (as described in chapter 2): group 1 phones are considered capable of realisations involving onset, core and offset; group 2 phones, defined as they are as being terminated at the beginning of the glide phase, are initialised to onset and core subphones only; group 3 phones, defined as beginning only at the termination of a glide element, are initialised to core and offset only. In all these three groups, however, the same principle is used for determining the initial subphonic analysis: the most stable portion is taken to constitute the phone's core, what precedes it (where applicable) is taken to constitute the onset, and what follows (ditto) to constitute the offset.

In reality, of course, a group1 phone may have no steady state, while after [-VOICE] phones, a phone from group 1 or group 2 may have no onset (conceived as entailing a period of formant-movement). The former fact is ignored in the initialisation, while the second is accommodated only to a limited degree, but it should be recognised that the goal of the initialisation is to get the allocation of vectors to subphones *approximately* right, or right *in as many instances as possible*, on the assumption that if this can be achieved the subsequent iterative resegmentation will remove the errors. Thus it is thought unlikely to be crucial that phones with no steady state will nevertheless contribute vectors to a core subphone in the initialisation, so long as in each class the majority or at least a large proportion of tokens do have steady states, so that the class sample comes in each case to be dominated by steady-state vectors; if this is achieved, vectors wrongly initialised to core because the most stable portion of their token was not stable in any absolute sense will be reallocated to onset or offset as appropriate subsequently. More misgivings arise over phones which begin in a steady state, for reasons which will be discussed further below.

The most stable part of a phone (for phones in groups 1 to 3) is found by first identifying its most stable quarter (discounting the initial and final vectors, which are automatically assigned to onset and offset respectively). In measuring

acoustic stability to identify the most stable quarter, the dynamic coefficients and power are ignored, only the first ten cepstral coefficients being taken into account. To identify the most stable quarter of the phone, the degree of correlation is taken between each pair of contiguous frames, and the cumulative correlation is then calculated for all sub-sequences of frames that constitute a fraction equal as far as possible to one quarter of the phone; the sub-sequence with the greatest cumulative correlation is taken as the most stable quarter of the phone. The degree of correlation between $x$ and $y$ is defined as

$$cor(\mathbf{x},\mathbf{y}) = \frac{(\mathbf{x},\mathbf{y})}{\| \mathbf{x} \| \| \mathbf{y} \|}, \tag{4.12}$$

where $(\mathbf{x},\mathbf{y})$ is the scalar product of $\mathbf{x}$ and $\mathbf{y}$ and $\| \mathbf{x} \|$ is the Euclidean norm of $\mathbf{x}$. [7]

The most stable quarter may be located anywhere within the confines of a phone (or more accurately, within the confines of a phone minus its initial and final vectors (for group1 phones) or minus its initial vector (for group2 phones) or minus its final vector (for group3 phones)). The most stable quarter may abut the initial vector or the final vector, or there may be unassigned vectors between it and the initial vector and/or between it and the final vector. The next step in the initialisation process is therefore to allocate the unassigned vectors to appropriate subphones. Intuitively, it seems that a solution to this problem should be based on whether an unassigned vector appears to have more in common with the core than the onset, or more in common with the core than the offset. In trying to measure this formally, a vital consideration is the massive overlap between contiguous frames; each vector represents an analysis of 32 ms of speech, but as there is only a 6 ms advance from one frame to the next, two contiguous frames share 26 ms of data, and the corresponding vectors of course reflect this. The problems this may cause may be illustrated by considering a phone with no real

[7] This measure is also referred to as the direction cosine similarity measure, since it measures the angle between the vectors, a value of 1 representing perfect match and a value of 0 indicating that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to each other. The measure is equivalent to a Euclidean distance measure for vectors whose elements have been normalised. The scalar product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is found by first multiplying each element of $\mathbf{x}$ by the corresponding element of $\mathbf{y}$, and then summing the products. The Euclidean norm of a vector $\mathbf{x}$ is equal to the positive square root of the scalar product of $\mathbf{x}$ with itself.

onset (e.g. a stressed vowel following an aspirated [-VOICE] stop burst). In such a case, the stretch of speech identified as the most stable quarter is likely to outscore other possible stretches by only small margins, the whole phone prior to offset being relatively stable; in these circumstances we would really like all or any vectors preceding the most stable quarter to be added to the core (leaving only the initial vector as a nominal onset). On the other hand, where there are real formant-transitions we would like the onset to end where the transition gives way to the steady state (should there be one). The procedure adopted reflects a greater concern with the avoidance of building up spurious onsets in the former sorts of case, than with optimal boundary-location in the latter sorts of case, in that it makes it probable that in cases with real formant-transitions, the later part of the transition may be initialised to the core, with the subsequent iterative process being relied upon for the required boundary-adjustment.

The rest of the initialisation, then, proceeds as follows (for phones in groups 1 to 3). (I give here the case for the onset vs core decision; that for core vs offset follows analogously, working outwards from the first post-core vector towards the end of the phone). Working outward from the vector **x** immediately preceding the initial core, we compare **x**'s correlation with (1) the mean of the initial core vectors and (2) the first vector of the phone (already assigned automatically to onset); if the correlation is greater with the core, we assign **x** to core and repeat the process for the next vector (working outwards, back toward the beginning of the phone), still using the mean of the original core for the comparison, without any updating. As soon as we reach a vector **x** that correlates more closely with the initial vector of the phone (should we do so at all), we assign **x** and any vectors appearing earlier than it to onset.

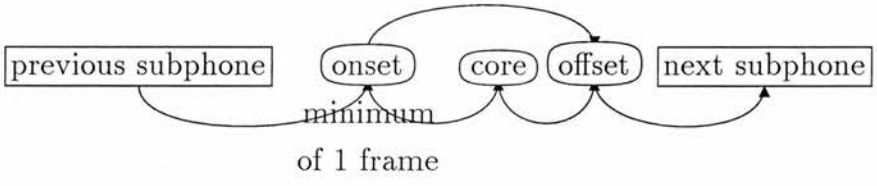
Little has been said so far about group 4 and group 5 phones. Glide elements of three-phase diphthongs that have no 'D1' element are treated in the same way as the consonants in group 5: all are initialised simply by splitting the phone in half as nearly as possible, and assigning the left half to onset and the right to offset. Diphthongal glides that follow cognate 'D1' elements are treated as unitary elements (not divided into subphones) for reasons which will be discussed in Chapter 6.

### 4.5.2 Iterative Closed-Test Reclassification

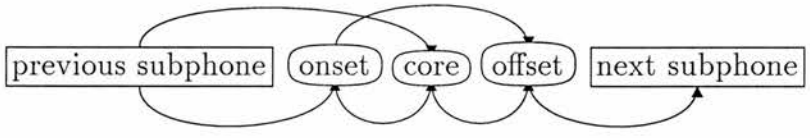
All phones initialised to subphones are given a 'resegmentation code' which determines certain limits set on the subsequent resegmentation process. The codes assign the phones to one or other of the following categories (once again the reader may refer back to chapter 2 for full details of the reasons behind the different treatments of different categories of phone):

1. monophthongal vowels and [y] following [-VOICE] phones, and 'D1' elements of three-phase diphthongs following [-VOICE] phones when they are not followed by a cognate diphthongal glide: all these phones are constrained to have at least an initial vector allocated to onset, until the final iteration when disappearance of an onset subphone is allowed; core states are optional, except when the onset disappears; all phones are constrained to end in at least one frame allocated to offset:

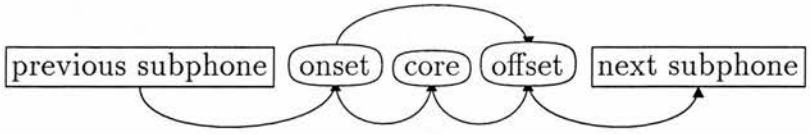
   (a) non-final iterations –




   (b) final iteration –

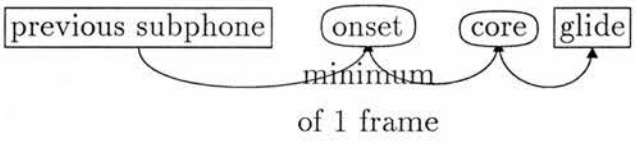



2. monophthongal vowels and [y] following [+VOICE] phones, and 'D1' elements of three-phase diphthongs following [+VOICE] phones when they are not followed by a cognate diphthongal glide: these are constrained throughout to begin in onset; cores are optional as with the first group, and conditions on occurrence of offsets are again as in the case of the preceding group.

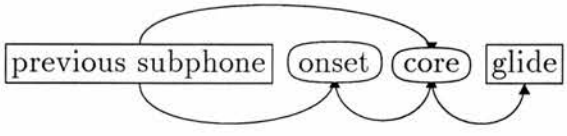



3. 'D1' elements of three-phase diphthongs following [-VOICE] phones and preceding a cognate diphthongal glide: these phones are constrained to begin with at least one frame in onset, until the final iteration when the onset may disappear; they are constrained throughout to end in at least one frame assigned to core.

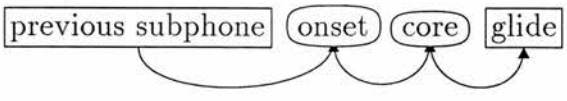
   (a) non-final iterations –



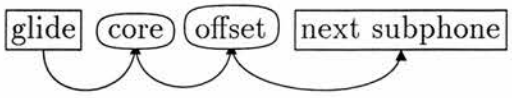

   (b) final iteration –



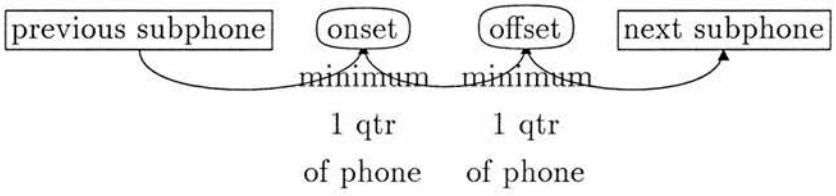

4. 'D1' elements of three-phase diphthongs following [+VOICE] phones and preceding a cognate diphthongal glide: these phones follow the conditions for the preceding group except that the constraint on beginning with at least one vector assigned to onset is maintained throughout.




5. 'D3' elements of three-phase diphthongs: these are constrained to begin with at least one vector assigned to core and to end with at least one vector assigned to offset.

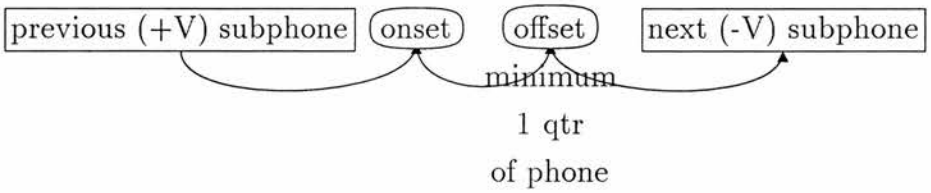



6. fricatives, affricate-releases, non-dark laterals, nasals and [h], apart from [+VOICE] fricatives and affricate-releases appearing between preceding [+VOICE] and following [-VOICE] phones: the initial quarter of the phone is reserved throughout for onset, and the final quarter for offset, so that at most three-quarters of any one of these phones can be allocated to one or other subphone.

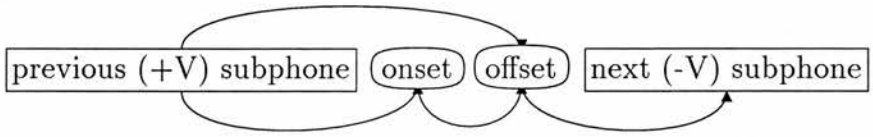



7. Lenis fricatives and [jhb] (frication phase of the affricate [jh] as in "Joe") appearing between a preceding [+VOICE] and a following [-VOICE] phone: the final quarter is reserved throughout for the offset subphone, but only the initial vector is reserved for the onset, and that only until the final iteration, when allocation of the complete phone to 'offset' is allowed. (The reasoning behind this was given in the subsection on fricatives in section 2.4.2.)

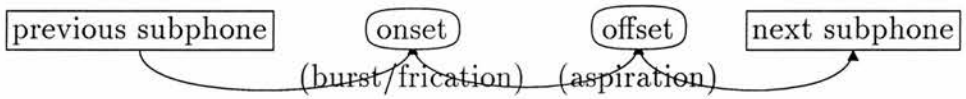
   (a) non-final iterations –



(b) final iteration –

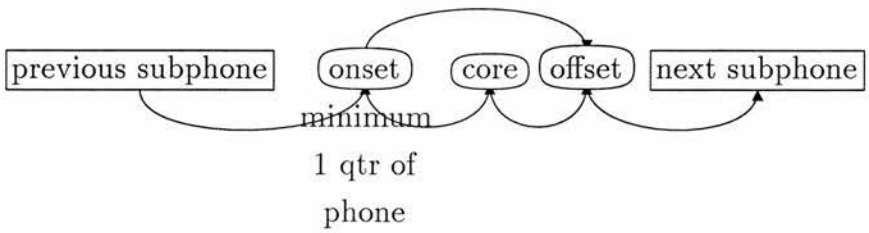



8. prevocalic aspirated [-VOICE] stop-releases: the first vector is reserved throughout for the 'onset' (burst-frication), and the last vector is reserved throughout for 'offset' (chiefly aspirative phase). (The reader may need to be reminded that other stop-bursts are not subjected to subphonic analysis.)



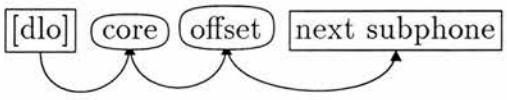

9. Dark [l]: when not including an initial [dlo] subphone,[8] this phone has its first quarter reserved for onset; a core state is optional, and the phone is constrained to end with at least one vector assigned to offset. When a [dl] follows a [dlo] (a devoiced onset subphone of a dark [l]), it is analysed in terms of core and offset only, and constrained to begin with at least one frame in core and to end with at least one vector assigned to offset.

(a) without initial [dlo] subphone –




[8]The 'dlo' label is used to designat an initial voiceless phase of a dark lateral.

(b) following [dlo] subphone –




## 4.6 Further details of the Recognition Procedure

### 4.6.1 Considerations of Efficiency

The most important factor in recognition is obviously accuracy but because of the complexity of the problem efficiency is also a major consideration. Using the methods of 4.3 above for scoring class-membership, and those of 4.4 for tracing the best utterance-level path, a possible approach would be to perform the two tasks in sequence: first compute the scores for each class at each frame, and then trace the best path. Working with full covariance matrices (not assuming complete statistical independence of the elements comprised in the representation vectors), each calculation of class-membership involves at best 152 multiplications, so that with 750 classes and a 1000 frame utterance, this approach involves us in 114 million multiplications just to get the class scores, before any search for the best path.

A possible (but again uneconomic) approach to the DP search, given the detailed sequence-constraints made possible by the definition of the phonetic classes, would be as follows: at each frame $j$, for each class $i, j$, go through the classes $k, j-1$, and if any $k, j-1$ is a legal predecessor of $i, j$, consider the partial path-score up to $k, j-1$ as a candidate for best path to extend to $i, j$; do this for all $k, j-1$ that are legal predecessors of $i, j$ until the best-scoring one is found, and extend that path to $i, j$ as the best path to it (to $i, j$). Record the index of the best $k, j-1$ (and the score of the newly extended path to $i, j$). Do this for all $i, j$, and at all $j$ up to and including the final frame of the utterance. This approach involves looking up a table or other information-source to see if $k, j-1$ is a legal predecessor of $i, j$, and doing this for each $i, j$, involving $750^2 \times 1000 = 562,500,000$ lookups for our example case.

Neither of these two approaches, of course, is optimal in terms of efficiency. It makes sense to allow for the abandonment of paths which become highly improbable, and it follows that once a path to any $i, j-1$ is abandoned, the score of any class $i$ at $j$ that is linked *only* to class $i$ at $j-1$ becomes irrelevant, so that it makes no sense to compute it. Given the way dynamic programming works, it is possible to do the scoring and the searching concurrently (or very nearly so): having traced the best-scoring path to each $k, j-1$ at the most recently completed stage, and pruned paths that seem the most improbable, we score an $i, j$ only if it has at least one live legal predecessor in the form of a $k, j-1$ that ends a non-pruned best path to frame $j-1$.

One method for pruning unpromising paths is by beam search (Lee & Hon 1988; Bisiani *et al.* 1989). A constant beam width $W$ is determined beforehand by trial and error (one which maximises efficiency without threatening to exclude the correct path), and applied as follows: at the completion of score-calculations at any stage $j$ in the search, the best path overall is identified out of all the best paths to the set of $i$ at $j$, and of the remaining paths, only those with scores within $W$ of the overall best path's score are saved, the others dropping out of contention. Thus while $W$ itself is a constant, the pruning process adapts its severity to the data somewhat, in that at each $j$ one edge of the beam (so to speak) is determined by the best overall score to frame $j$, so that if all scores are relatively poor at $j$, the upper edge of the beam will fall somewhat higher.

As far as the look-up of sequence-constraints is concerned, it is of course the case that the same set of constraints holds at every $j$, given a fixed ordering of the classes. It is therefore possible to define once for all the indices of classes that need to be considered for any $i$ at any $j$. Instead of having to work our way through *all* $k, j-1$, and check for each one if it is a legal predecessor of the class $i$ of current interest at $j$, we work automatically through a pre-prepared list of just the classes that *are* legal predecessors of $i, j$, regardless of what $j$ is. No look-ups at all are required.

Further efficiencies are possible. Firstly, evaluation of the term

$$-0.5(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i)$$

of (4.9) can be simplified by taking advantage of the symmetry of $\Sigma^{-1}$. [9] If the diagonal terms in $\Sigma^{-1}$ are halved at the time $\Sigma^{-1}$ is calculated during training, and the 0.5 factor is dropped from (4.9), we can reduce the computation by almost half by working with just the upper or lower triangle of $\Sigma^{-1}$ (including the diagonal).[10]

For a 16-dimensional vector, we reduce the number of multiplications from 272 to 152, and since the evaluation of (4.9) is a major cost in the recognition procedure, this is a very valuable saving. (The figure of 114 million multiplications given earlier for the example case using inefficient methods would actually be 204 million without this saving).

A second further efficiency can be achieved via a dual scoring process which begins with a quick preliminary calculation of the probability of class-membership, and proceeds to more careful scoring only if the preliminary score is below a certain threshold (otherwise returning that threshold as the class-membership score). The rationale for this is as follows: accurate evaluation of scores is important only for classes which have some significant probability of being the correct class, and since accurate scoring is computationally expensive, it makes sense to sacrifice accuracy for those classes which do not, from first inspection, appear to be likely contenders. One method for implementing this efficiency is to use an assumption of statistical independence of vector elements for the preliminary

[9] I am indebted here to Stephen and Michael Isard, both for the fact and the correct encoding of it.

[10] The normal sequence of row by column multiplications is truncated, if we work with the lower triangle of $\Sigma^{-1}$, by progressively dropping a further term from $(\mathbf{x}-\mu_i)^t$ for each product, so that we calculate

$$(x_1 - \mu_1, x_2 - \mu_2, \ldots, x_N - \mu_N) \times \Sigma^{-1}_{1,1toN}$$

$$(x_2 - \mu_2, x_3 - \mu_3, \ldots, x_N - \mu_N) \times \Sigma^{-1}_{2,2toN}$$

$$\vdots$$

$$(x_N - \mu_N) \times \Sigma^{-1}_{N,N}.$$

scoring, and take just a subset (say, the first six) of the elements of the representation vector, first scoring using just the subset and the diagonal matrix. It should be noted that in this scheme no class is given a *worse* score than it gets via the preliminary scoring; it should also be noted that the cepstral coefficients are ordered more or less according to their variance, those with the greater variance coming first, with approximately 85% of the global variance being accounted for by the first 6 coefficients. Tests on a very small number of utterances revealed *no difference* in the output transcription whether using preliminary scoring of this kind or not, while the saving in computation-time was considerable.

Further efficiencies are obtainable by converting to fixed point arithmetic once a score has been calculated, by reducing the frame-rate to a minimum (e.g. to 100 per second instead of the 160 or so per second used here), and perhaps by first identifying frames which come within a certain measure of similarity and coalescing them (this last technique has at least been used in HMM-based systems employing vector quantisation (e.g. (Chow *et al.* 1986)) and found not to affect performance). (A conversion to fixed-point arithmetic is effected in the transcription program used in FURIDA, all the original scores being multiplied by 10 before the conversion to reduce the loss of accuracy.)

Further economies are possible, but some of the commonly used ones involve making assumptions about the data which are in fact difficult to justify. They will be looked at under a different heading in Chapter 5 (5.5).

### 4.6.2 TR classes in the Transcription Procedure

The transcription (scoring and searching) process is made considerably more complicated by the inclusion of TR classes, for two reasons: firstly, as described in 3.5 TR classes are built in training from sequences of at least two and at most three vectors around frame-boundaries, and it is therefore appropriate to enforce the same minimum and maximum duration for sequences of TR class frames in recognition; this obviously involves keeping counts of TR frames in paths and consulting such counts as appropriate. Secondly, whereas in cases not involving TR classes legal sequencing between $i, j$ and $k, j-1$ is all that needs to be considered when determining whether the two may link in a path, where

a sequence of TR frames occurs we also need to consider linkage across the TR interval, i.e. between classes linking up with the TR sequence at each end. It will be convenient to refer to such linkage across a sequence of TR frames as cross-linkage or remote linkage.

In an early "solution" to the problem of enforcing remote sequence constraints, (using the example in figure 4.2), once $k, j-3$ had been identified as the best predecessor of TR $t, j-2$, none of the classes $i$ at $j+1$ which had TR $t, j$ as their best predecessor could actually continue live paths unless they were also cross-linked to $k, j-3$. Further thought led to the realisation that this strategy ran counter to the spirit of DP, which requires that at each stage $j$ each $i$ should (pruning of unpromising paths apart) have a best path to itself. With the strategy just described, however, (continuing with the example) many classes $i$ at $j+1$ are prevented from continuing live paths simply because of the selection of a single best predecessor to TR $t, j-2$; it must be remembered that the TR classes are highly generalised (TR1, e.g., covers borders between any vowel, [r],[w], [y] or [dl] on the left side and any stop-closure (and a few other categories besides) on the right), so that the short-termism implicit in this strategy could actually cut out a lot of potentially promising paths.

A somewhat different strategy was therefore adopted. Continuing with the example, for each $i$ at $j+1$ that has TR $t, j$ as its best predecessor, we search at $j-3$ for the $k$ that is the best "remote predecessor", and record that for future reference. Thus in general, any non-TR $i, j$ that has a TR $t, j-1$ as its best predecessor will also have a remote predecessor, and both are recorded for eventual backtracking. For backtracking to be possible, we need to know the length (2 or 3) of any TR frame sequence, but this information is already available, having been required during the DP search as described earlier. Thus whenever in backtracking we find that for a non-TR $i$ the best predecessor is a TR, we look up the length of the TR sequence, look up $i$'s best remote predecessor, and shift to the latter the appropriate number of frames further back, taking up backtracking again from there.

This strategy has the benefit of allowing paths which are trailing the best at some point $j$ to catch up over the slightly longer term. This may be seen by contrasting the old "solution" with the new one when applied to the following

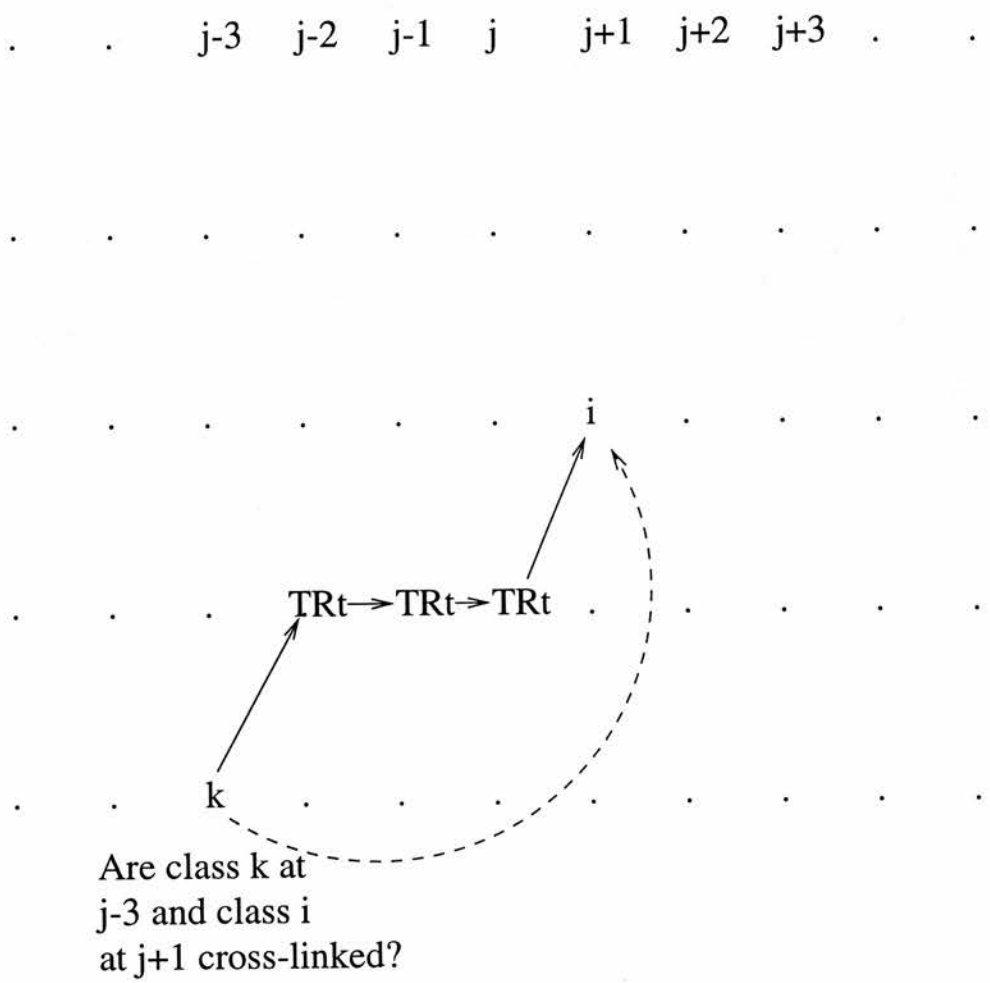



**Figure 4.2.** Linkage Across TR 'Segments'

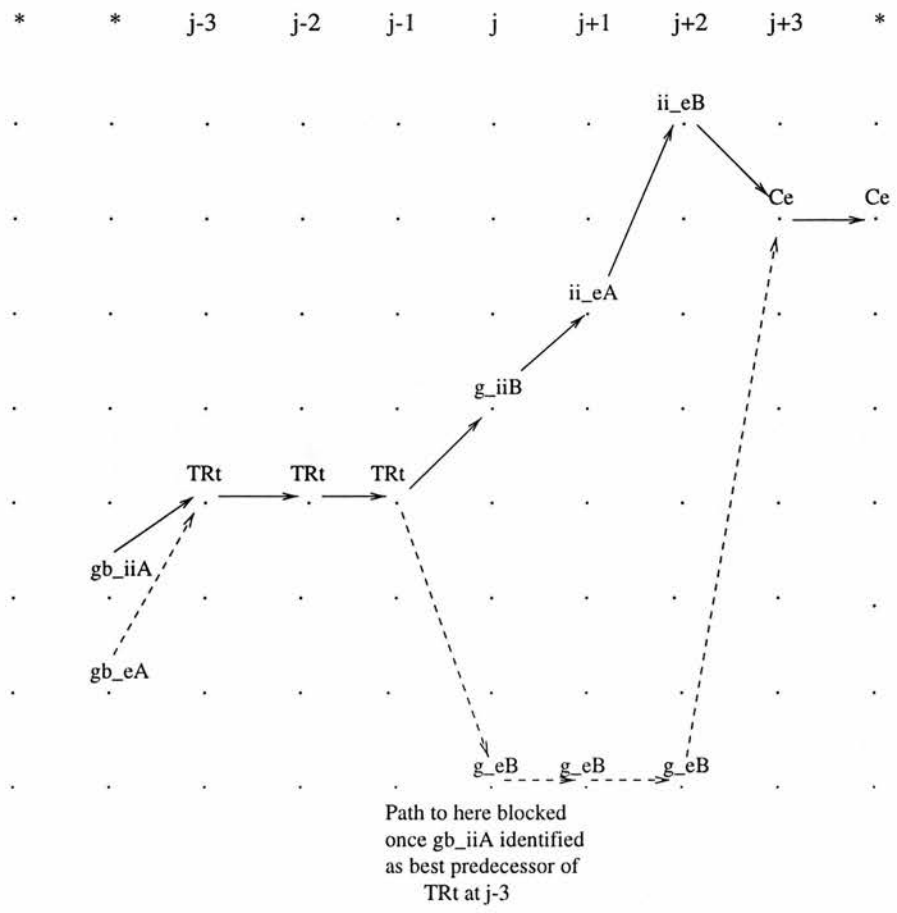



**Figure 4.3.** Old Scheme for Remote Sequence Constraints

hypothetical (but realistic) situation: imagine we have as part of an utterance an [e] vowel preceded by a [gb]. The "velar pinch" typical of the earliest part of the vowel tends to give rise to insertions of [ii]. Under the old system there was no chance at all of recovering from such an insertion, while under the later system there is still a possibility of the [gb e] path catching up, if only the [gb_eA] class scores sufficiently more highly than the [gb_iiA] class to cancel out the earlier deficit (figures 4.3 and 4.4).

Having found the best remote predecessor for each $i, j$ that has a TR $t, j-1$ as best predecessor, we set the score for the path up to $i, j$ (inclusive) equal to the best path score to the best remote $k$ plus the individual scores of the two or

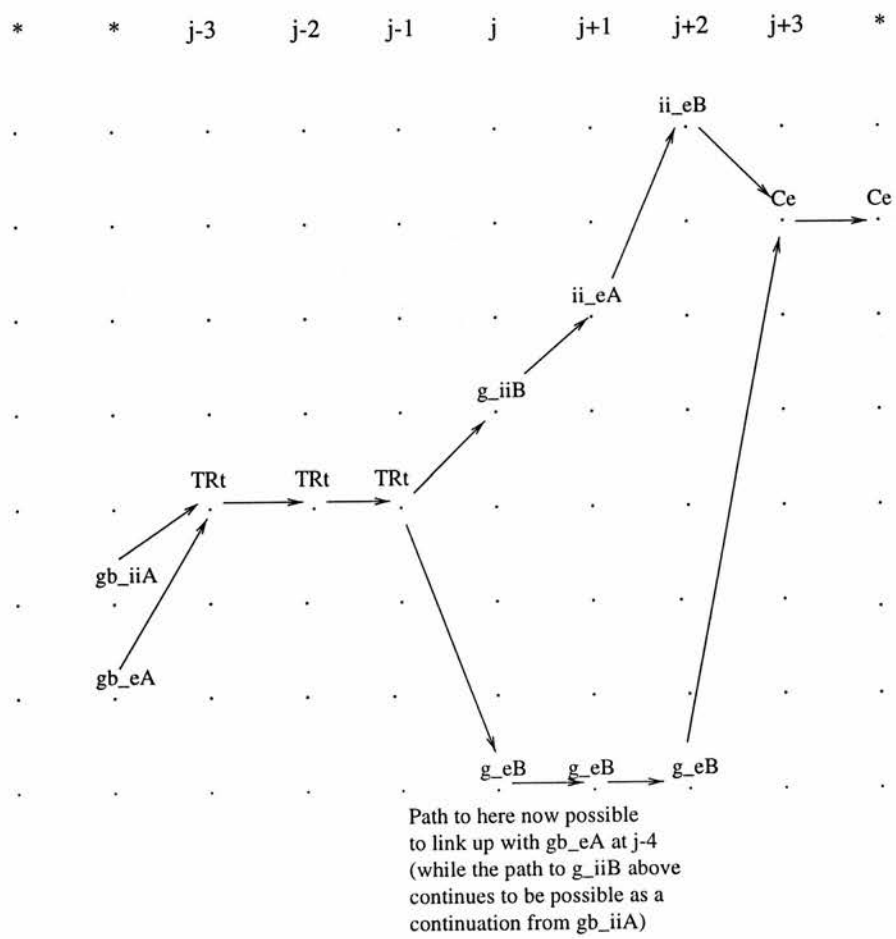



**Figure 4.4.** New Scheme for Remote Sequence Constraints

three TR class frames plus the individual score of the class $i, j$ concerned. Note that in conducting the search for best remote predecessor of $i, j$, we insist not only that the candidates be legal predecessors of $i, j$, (and distinct from $i, j$ — recall that we make no distinction between geminates and single phones), but also that they be (non-TR) legal predecessors of the TR $t, j-1$ that is the best immediate predecessor of $i, j$.

It should be noted[11] that the implementation of TR-class duration-limits and remote sequence-constraints just described does not guarantee that the best path will be found. Firstly, as regards the precise location of TR-class 'segments': because of the enforcement of a minimum and maximum duration, once a TR $t$ at $j$ is found to have TR $t$ at $j-1$ as its best predecessor, the latter being the last of the series of TR $t$'s, the option of having rather the TR $t$ at $j$ as the first of the series, with further TR $t$'s at $j+1$ and at $j+2$, is precluded, a TR $t$ at $j+2$ being ruled out by the maximum duration limit; yet the location at $j$, $j+1$ and $j+2$ might well have been the better one (compared with that taken at $j-1$, $j$ and $j+1$). This problem can be solved by modelling each TR class with two or three single-element classes — for any TR $t$ in the earlier formulation, we would have three independent classes, say TR $t^{(1)}$, TR $t^{(2)}$ and TR $t^{(3)}$, and so could allow the optimal location to be found for the TR "segment" by keeping separate scores using the independent elements, as illustrated in figure 4.5.

There would appear to be no objection to using the single set of class-parameters ($\mu, \mathbf{\Sigma}$ for TR $t$) for all of the three elements, should data-shortages make this necessary (duplicating the single set of parameters to produce one copy for each set — compare a similar ploy for stop-closures described in 4.6.3).

The use of independent elements would appear to offer a solution to a further, related problem which is connected with the fact that TR classes are normally highly generalised, with a great many specific 'interpretations' (thus borders between [e] and [VSAcv], [w] and [NVSAc], [iiD3] and [VNVSAcv], and many more, are all represented by the single TR class TR1). Without single-frame decomposition of the kind just envisaged, it is possible that the optimal placement of a TR 'segment' could vary for different interpretations (in different paths, one

[11]This and the succeeding paragraph lean heavily on comments and suggestions from Fergus McInnes, who pointed out both the problem and the solution described here

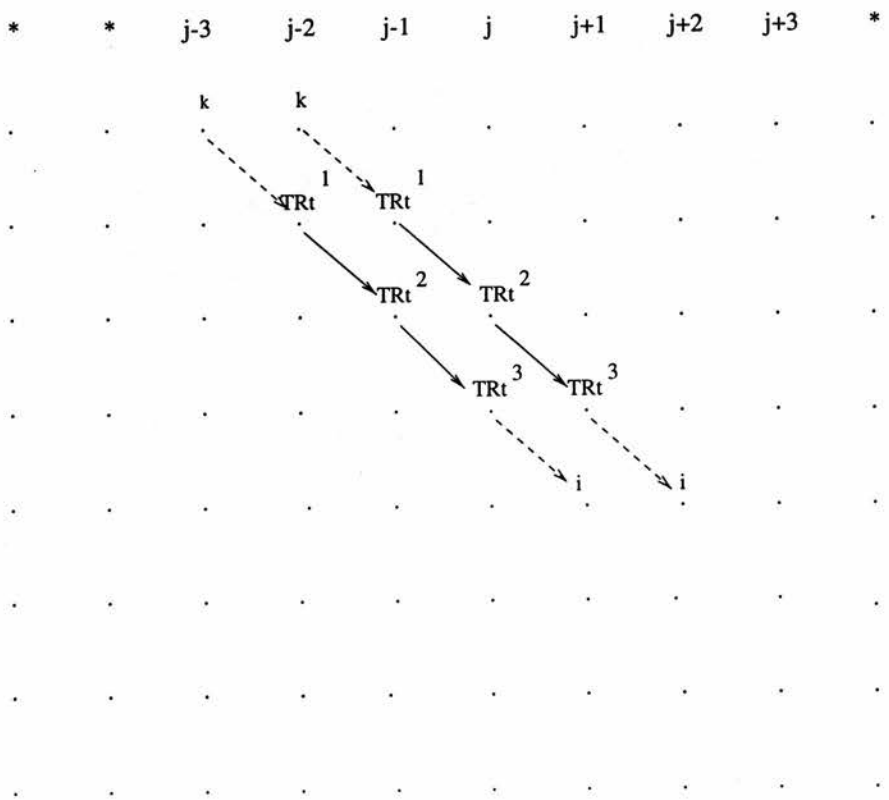



**Figure 4.5.** Allowing Optimal Placement of TR 'Segments' (I)

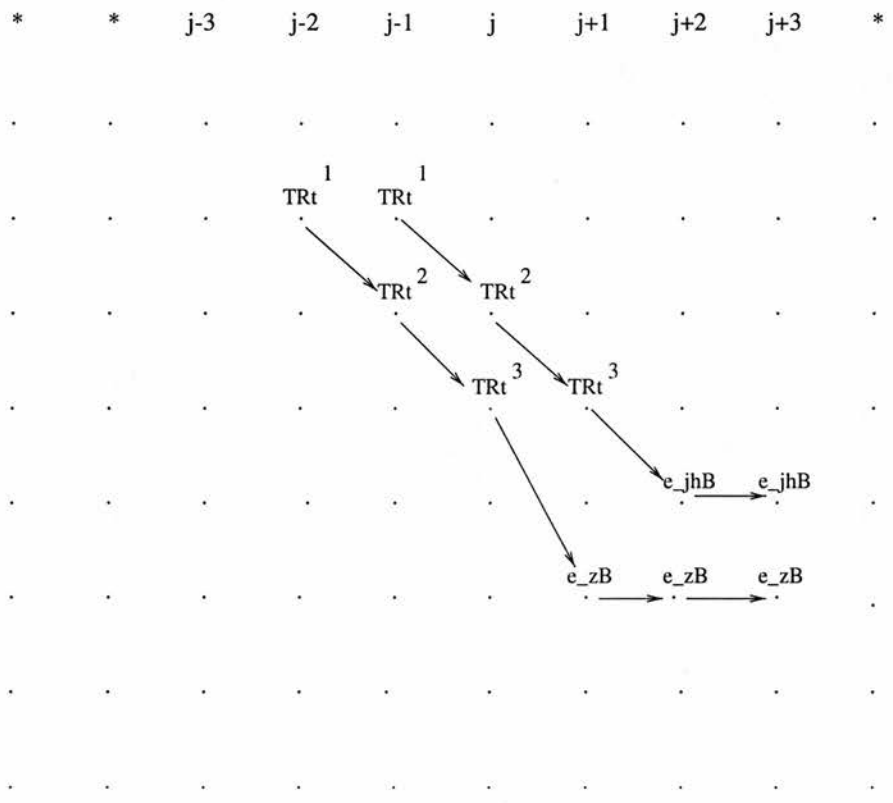



**Figure 4.6.** Allowing Optimal Placement of TR 'Segments' (II)

could equally well say); yet in the implementation described above, once a TR segment begins to be consolidated, its placement is determined for *all* possible interpretations. Given decomposition into independent elements, however, it would appear to be possible to allow for variable placement for different interpretations, as illustrated in figure 4.6.

The fact that there may be large numbers of specific interpretations of a particular TR class would not appear to present an obstacle, since having completed a $TRt^{(1)}$, $TRt^{(2)}$, $TRt^{(3)}$ sequence in a given path, the remote search then carried out requires to find only the single best remote continuation, and there would now of course be a number of different placements of the given TR class 'segment', each having a remote continuation found by remote search.

Results of decomposition of TR classes into independent elements are described in chapter 6 (6.2 and 6.7).

### 4.6.3 Stop Closures in the Transcription Procedure

As explained in Chapter 2, 2.5.2, lack of acoustic distinctiveness between many individual stop-closures prompted the use of generic classes, namely NVSAc, VSAcv, VSAcnv, NVVSAc, and VNVSAcv (where the 'v' or 'nv' termination points to the VOICE feature of the phone preceding the closure). The original idea was to rest content with an automatic transcription in terms of such generic labels, but the goal of recovering the specific identities from context was too appealing to ignore, and in this section I describe the strategy employed for attempting to achieve this.

If it were possible to know a priori whether a particular stop-closure was associated with a single rather than with a double stop, it would be a fairly simple matter to assign a specific identity to it (in DP, the interpretation would have to be consistent with the contextual annotation of the class of the best predecessor, and would then determine which subphones could continue the path in question; for a [tc], for example, only phones with offsets compatible with following [tc] could precede, and only versions of [tb], or subphones capable of following immediately after an unreleased [t], could link up with it subsequently). However, generics like NVSAc, VSAcv and VSAcnv may denote either single or double stops (2.5.2), and in such cases we have no idea whether compatibility should be enforced between the phones to either side of the closure.

Given that we do not know beforehand whether a single stop or a pair of stops is involved, why should we not simply let the burst (or if this is absent, other contextual information given via the following phone's label), and the transition into the closure, determine the interpretation of the generic label, allowing whatever gives the highest score to be the winner as everywhere else? Thus if the best path to the closure ends with a phone whose right-context annotation is for a [d], and the best path from the subphone that follows the closure begins with a version of [kb], why not just interpret the closure as a [dkc]? The first point to be made in response to this is that the generics do have their own statistical parameters, and

that the scores for the closures themselves therefore do have some influence on the proceedings. So we can't entirely go along with the proposal just aired; the generic appropriate to [dkc], namely [VNVSAcv], might score much worse than [NVSAc], for example, and so push the decision in a different direction. But why not let this other interpretation win, whatever it may be? The obvious answer to this is that we can't altogether ignore duration either, even if the question of the proper weight to give it is a difficult one to determine. Intuitively, it seems reasonable to suppose that below a certain duration the likelihood of a double stop must become smaller, while above a certain duration the likelihood of a single stop must do so.

It is in fact one of the central contentions of this thesis that since duration is not primarily a phonetic phenomenon (not primarily determined by phonological identities of phones), it must always be dangerous to use duration-statistics gathered without regard to higher-level factors to help determine phonetic identities. More will be said on this in sections 4.7 and 4.8. It was, however, decided to accept a role for (pseudo-) duration-probabilities in the determination of specific identities of generic stop-closures, while keeping their influence on events small. Histograms for single and double non-glottalised closures were built from the training-data and made the basis of penalty-assignments to stop-closure frame-sequences. The penalties themselves are as shown in table 4.3. (Note that the penalties shown are multiplied by 10 when used in the transcription algorithm, to harmonise with the scaling up of the spectral probabilities.)

(If some of the values for short-duration double stops seem surprising, recall that TR frames will take up some of the closure at each end — on average, one might suppose, three frames' worth of the duration of any closure will be taken up in this way. Note that the values of 23 represent negative log values of $1E-10$ used where the original histogram values are actually zero.)

For durations of 8 to 11 frames the histogram counts (or relative frequencies derived from them) were in fact equal; in order to force a decision between pairs like [tc] and [ttc], 1 was added to the penalties for double stops for durations of 8 to 11 frames, on the basis of (assumed) greater a priori probability for single as compared with double stop-closures. All penalties are rounded off to the nearest integer, after being derived as negative logs of the smoothed histogram values (in

| Stop Penalties | | |
|---|---|---|
| No. of frames | Single Stop | Double Stop |
| 1 | 2 | 23 |
| 2 | 2 | 4 |
| 3 | 2 | 3 |
| 4 | 2 | 3 |
| 5 | 2 | 3 |
| 6 | 2 | 3 |
| 7 | 2 | 3 |
| 8 | 3 | 4 |
| 9 | 3 | 4 |
| 10 | 3 | 4 |
| 11 | 3 | 4 |
| 12 | 4 | 3 |
| 13 | 4 | 3 |
| 14 | 4 | 3 |
| 15 | 5 | 3 |
| 16 | 5 | 3 |
| 17 | 6 | 3 |
| 18 | 6 | 3 |
| 19 | 7 | 3 |
| 20 | 7 | 3 |
| 21 | 8 | 3 |
| 22 | 8 | 4 |
| 23 | 9 | 4 |
| 24 | 23 | 23 |
| 25 or more | 23 | 23 |

**Table 4.3.** Stop Duration Penalties

turn expressed as relative frequencies); the rounding off is effected because the DP procedure is carried out using integer arithmetic. Duplication of computational work was avoided by calculating scores just for the original generic classes, and then propagating the scores to all of the appropriate specific forms.

In DP search, then, the best path to, for example, a [tc] and to a [ttc] at column $j$ will normally be the same initially [12], but on exiting the stop-closure sequences (via a TR class), duration-penalties are applied, so that it is the duration-penalties themselves that come to determine the choice between geminate and non-geminate pairs like [ttc] and [tc].

Implementation is made easier by the presence of TR classes, since a path never *enters* a stop closure sequence except from a TR class, and never *leaves* a stop closure sequence except via a TR class. This provides simple criteria for starting and ending counts of closure frames during DP search: if class $i, j$ is a closure and its best $k, j-1$ is a TR class we start counting from scratch; if $i, j$ is a closure and its best $k, j-1$ is the identical closure class, we increment the count; and if $i, j$ is a TR class and its best $k, j-1$ is a closure class we call the count to an end and apply an appropriate penalty. The penalty will subsequently be taken into account when we do a remote search from the "other side" of the TR class interval for the best remote linkup there (figure 4.7).

How significant is the role of duration-penalties in affecting the overall probability of a switch in PLACE and/or VOICE across the closure? For most durations the differences between penalties are small (for durations between 2 and 14 frames the difference is only 1 point), so that it is unlikely that penalties will be able to influence decisions greatly when the choice is not simply one between geminate and non-geminate pairs like [tc] and [ttc], since spectral characteristics of the burst, and of the transition into the closure, are likely to affect scores much more dramatically, while in some cases spectral characteristics of closure-frames themselves will sway things one way or another too. In some cases, of course, one may find oneself wishing for a scheme in which closure-durations were given

[12] Where the set of predecessors of a single stop-closure is not exactly the same as that of its geminate partner, the two classes may be arrived at by different routes and so (probably) have different scores; [gcv], for example, includes [s_gA] (offset of [s] before [g]) in its predecessor set, but [ggcv] does not, [s g g] not being a legal sequence in English

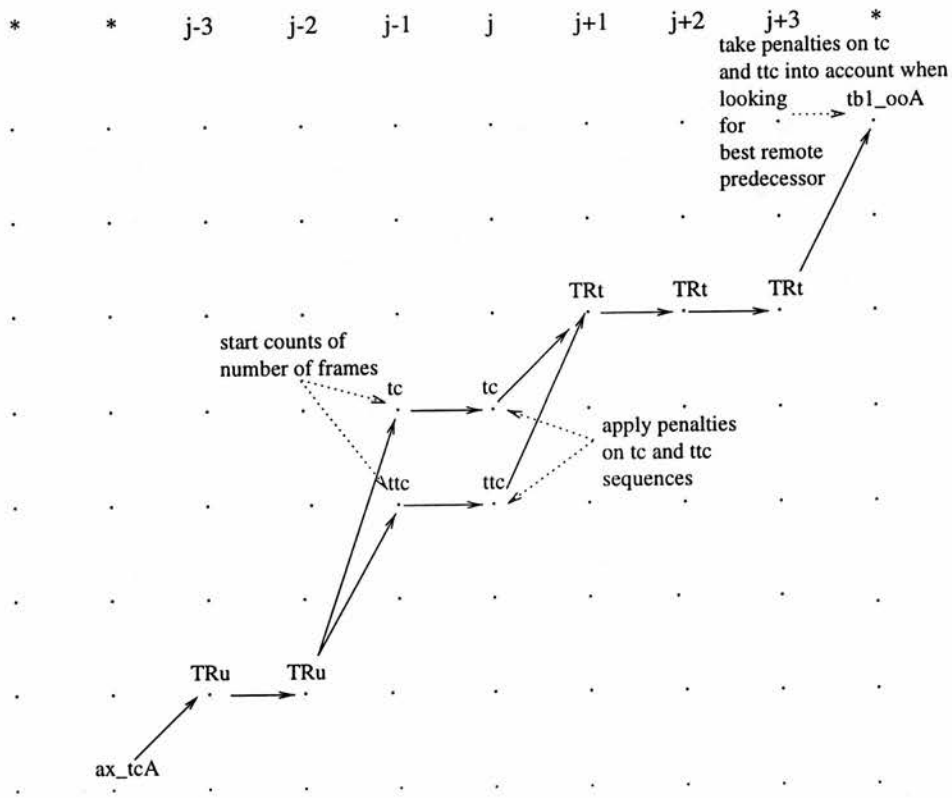



**Figure 4.7.** Application of Stop-closure Duration-Penalties

more weight, as when getting a transcription such as [x_gA (TR's) gbcv (TR's) bb_xA] with a duration for the closure of only two frames, since one would suspect that either the classification of the burst or that of the transition into the closure was wrong in such a case, and would have liked the duration-penalty to have been sufficiently great to swing things in the direction of a single stop. (The point made earlier about the brevity of some transitions into closures, particularly when TR classes are taken into account, is of obvious relevance here.) It is freely acknowledged here that the penalty-scheme adopted is something of an ad hoc and rather hastily executed response to a specific problem; the question of how to arrive at a scheme in which *appropriate* weight is given to all the factors involved in the interpretation of stop closures has not been seriously addressed at all here.

### 4.6.4 Converting from the Subphonic Transcription to A Phonetic Transcription

This is trivial for the most part. Recall that in subphone names ending with 'A' and containing an underscore, the part of the name preceding the underscore designates the phonetic (phone-level) identity, whereas in subphone-names ending in a "B" and containing an underscore, it is the part of the name between the underscore and the terminating "B" that designates the phonetic identity. The merely contextual information contained in the rest of the subphone-name can thus be stripped away, as can the "C" prefixed to core vowel, [y] and [dl] subphones, and where this results in identical consecutive designators, these are coalesced into a single label designating the phone:

```
AST: [ SIL SIL_sB s_aA s_aB Ca a_mA a_mB m_silA sil ]  --->

     [ SIL     s  s      a   a a      m  m      sil ]   --->

APT  [           s          a           m           ]
```

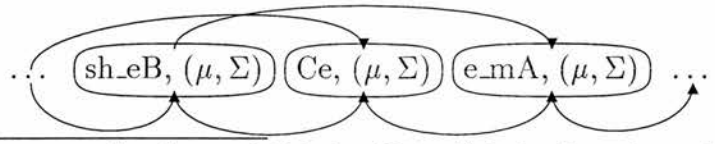
In some cases the conversion does not follow as simple a format as this, and individual rules need to be applied, as for example when dealing with 'context-independent' subphones like [mo] (a devoiced subphone of an [m]) or [NP] (a nasal plosion 'segment'), or with pairs of stop-burst subphones like [tb1_eA], [tb2_eA],[13] and so on, but the application of these rules is in almost every case straightforward.

The coalescing of identical consecutive symbols is safe given the policy of treating geminates as single phones in the first instance.

## 4.7 A Comparison with Hidden Markov Modelling (HMM)

### 4.7.1 Introduction

It will be clear from what has been said above that the attempt to transcribe speech in FURIDA is based on the use of statistical profiles $(\mu_i, \mathbf{\Sigma_i})$ of stages in the temporal development of phones. The temporal structure in that development — what stages are involved in the production of a phone, how they are sequenced, what stages if any may be omitted, and the like — is *stipulated* to be of such and such a form (or to fall within a specified range of possibilities) on the basis of what appears to be the case in general from examination of spectrographic records; the nature of the spectrographic data was described in some depth in Chapter 2. A schematic representation of how an entire phone is "modelled" (if the use of that word is justified) reveals a relatively liberal temporal framework, as this example for tokens of the vowel [e] occurring between [sh] and [m] demonstrates:




[13]The reader may need to be reminded that [tb1_eA] is the first phase of the release of a [t] preceding an [e] vowel (the phase typically dominated by a burst transient and frication), while [tb2_eA] is the second phase of such a [t], typically showing aspirative formants

The arrows indicate temporal sequencing, and show that in this particular case (a) the onset subphone is optional ([sh] being [-VOICE]) and (b) the core is optional following an onset (this being a trait that appears to be true in general of vowels, particularly the shorter vowels). (It may be worth reminding the reader, too, that this "model" is put together from unrelated bits of phonetic material, rather than from integral tokens of [e] appearing between [sh] and [m]: thus the statistical profile for [sh_eB] is drawn up on the basis of examples of [e] following [sh], regardless of what followed the [e], and that for [Ce] on the basis of [e] steady-states irrespective of phonetic context; this, however, is not an essential feature of FURIDA – given sufficient data, there would be no reason for not making the vowel's steady-state context-dependent (giving us, say, [sh_e_m] rather than [Ce]), and making the onset subphone sensitive to its remote right context [m] as well as to its immediate left context [sh], and similarly for the right subphone.)

In the most successful of standard approaches to ASR, Hidden Markov Modelling (Rabiner 1989), phones are modelled as unitary entities, and all the detail of the structure in their temporal development is *learnt* during training, though some framework is established a priori by giving models a definite number of *states* before training begins, and perhaps by stipulating that certain sequences of states will be possible, and others impossible. While the states of a hidden Markov model are essentially abstract, a useful starting-point for the newcomer to HMM is to think of these states as parallelling the subphonic elements of phones prominent in the present work. At least in this respect the parallel is a close one: the advantage of profiling subphones rather than entire phones is that one gets much sharper and therefore more discriminating profiles (certainly when working with single gaussians for each distribution) than one would if lumping the vectors from the entire phone into a single distribution, and a HMM is essentially a device or logical machine which makes it possible to achieve the same sort of resolution of a general collection of data into a set of sub-collections (each modelled by a state of the HMM), with precisely the same advantages in mind: the likelihood of a vector with respect to its own state or subphone will in general be greater than its likelihood would have been with respect to the phone as a whole, had the phone been modelled with a single lumped distribution. It should be stressed, however, that a HMM is in no way constrained to divide its

data into states corresponding to phonetic elements such as subphones attempt to represent.

A further difference between HMM and the technique espoused in the present work concerns the modelling of temporal structure. HMM has analogues of the connections between subphones indicated by the arrows in the figure for (sh)**e**(m) above, in the form of transitions between states. Each state will normally have "self-transitions" too (loops back to itself) – these are implicitly understood in FURIDA. In HMM all transitions, whether self-transitions or transitions from one state to a distinct state, have an associated probability learnt from training-data. Each advance from one frame to the next of speech involves consideration of which state of the model is most probable in the light of these transition-probabilities. Each state meanwhile also has associated with it a statistical profile which may take the same form $(\mu_i, \Sigma_i)$ as that used to model subphones in this work. The two sets of probabilities — transition-probabilities and spectral (or "output" or "emission") probabilities — are treated as independent of each other (so that scoring the fit between a stretch of acoustic material and a particular HMM involves multiplication of the probabilities in each of the two sequences, or addition in the log domain). In FURIDA, by contrast, the alignment of subphones with pattern-vectors is determined (in the basic implementation of the system) solely by reference to spectral probabilities: one progresses from an assignment to [sh_eB] to an assignment to [Ce] solely on the basis of the vector in question having a greater probability of belonging to [Ce], regardless of how great or small a number of vectors have already been assigned to [sh_eB] at that point. I shall revisit the issues involved in duration-modelling further below.

In HMM, then, speech is modelled as a dual process, involving an underlying Markov chain or succession of states,[14] and some form of statistical model for each state governing the likelihood of particular representation vectors appearing while in that state. This duality of process makes it difficult to form intuitions about how HMM actually works in phonetic terms – it seems likely that with a judicious pre-selection of numbers of states and of connections between states,

[14]To say that the state-sequence is Markovian is simply to say that the probability of any state being entered at time $t$ depends solely on the state that was occupied at time $t-1$; there is no memory beyond that.

the realities modelled will (for some large part of the time?) be interpretable in phonetic terms, but this cannot be relied upon.

The form of Bayes' Theorem applicable in the case of HMM makes reference to a *sequence* of vectors $\mathbf{X}$ rather than to a single vector $\mathbf{x}$. Using $\mathbf{M}$ to represent a particular HMM, we may state it as:

$$P(\mathbf{M} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{M})P(\mathbf{M})}{P(\mathbf{X})}. \tag{4.13}$$

This equation (which may be compared with equation 4.4) states that the probability of a particular HMM $\mathbf{M}$ given a vector sequence $\mathbf{X}$ is equal to the acoustic likelihood of the vector sequence given the model (the model's score for that sequence of vectors) times the prior probability of the model, divided by the probability of the vector sequence. Thus whereas in FURIDA the posterior probabilities are derived at frame level from the frame-level acoustic likelihoods (and class priors, in a realistic application), and the subphone and phone scores are arrived at by summing the (negative) logs of the posteriors, in HMM the acoustic likelihood ($P(\mathbf{X} \mid \mathbf{M})$) is itself derived by summing a sequence of frame-level (negative log) emission probabilities and transition probabilities. Perhaps the most significant feature of this difference is again the inclusion of transition probabilities in the calculation of $P(\mathbf{X} \mid \mathbf{M})$.

### 4.7.2 Training

Training in FURIDA is very simple: vectors are assigned to subphones in an initialisation motivated in large part by phonetic considerations,[15] and when this has been done for all phones in the training-data, means and covariances are estimated for each (generally subphonic) class, and the data is then subjected to closed-test reclassification using these estimates, the process of reestimating and resegmenting repeating until the estimates do not change. In the current implementation, resegmentation can take place only with respect to *subphonic* boundaries, that is, within the confines of a phone, the phone-boundaries created

[15]It is my hope that in those cases where the initialisation is simply a 50:50 division of a phone into two subphones, something more principled than the present crude limits could be found to constrain the subsequent resegmentation.

by the initial manual labelling being regarded as fixed. This is not an essential part of the system, however; this issue will be returned to below.

One of the methods available for training HMMs is known as *Viterbi* training, and is very similar to the method used in FURIDA, apart from the initialisation: in Viterbi training, the vectors of a phone are first divided *equally* among the states of the appropriate model, and when this has been done for all phones in the training-data, means and covariances for model-states, and transition-probabilities between states, are estimated. The training-data is then resegmented in the light of the initial parameters using the Viterbi algorithm, a form of dynamic programming which is essentially no different from that described earlier in this chapter, except for the additional feature of also taking account of probabilities on the connections between states when calculating subpath scores. New estimates are then derived from the new segmentation into states, and the data is segmented again in light of the new estimates, the process continuing until the estimates of means and covariances cease to change.

Viterbi training is based on finding the most probable path [16] through a model. *Baum-Welch* training is based rather on *total* probability, that is, the sum of the probabilities of all possible paths through the model. In this form of training, vectors are not assigned to particular states in an all-or-nothing manner, but each vector is given probabilities of belonging to all the states; one might say that a vector is assigned to a state to the extent that it is probable that the model would be in that state at the point in the temporal development of the phone represented by the vector in question. The Baum-Welch formulae for reestimating the means and covariances of a state of a HMM may be compared with the formulae given earlier for Maximum Likelihood estimation of the means and covariances of a subphone (equations 4.1 and 4.2); in the Baum-Welch formulae one finds an additional term (here, $L(t)$) representing the probability of being in the state in question at time $t$, and a normalising denominator, as well as a localisation to the context of individual states of individual phones (made explicit here by use of a time-related subscript $t$ attached to the vector variable):

[16] that is, sequence of states, where it must be remembered that 'sequence of states' may include repeated visits to "the same" state – the states are abstract.

$$\hat{\mu} = \frac{\sum_{t=1}^{T} L(t)\mathbf{x}_t}{\sum_{t=1}^{T} L(t)} \tag{4.14}$$

and

$$\hat{\Sigma} = \frac{\sum_{t=1}^{T} L(t)(\mathbf{x}_t - \hat{\mu})(\mathbf{x}_t - \hat{\mu})^t}{\sum_{t=1}^{T} L(t)}. \tag{4.15}$$

(I have refrained from making the dependence on a specific state explicit, in order to keep the formulae as simple as possible.) The probability of state occupation $L$ can be found efficiently using an algorithm known as the Forward-Backward algorithm (Rabiner 1989; Young *et al.* 1997).

One very attractive feature of HMM training is that it is possible to proceed without a phonetic labelling of the training-data, if one has an orthographic (word-level) transcription. This is of great practical benefit because phonetically labelling a given amount of training-data can take many, many times longer than simply typing it as text from audio. Training begins by consulting a pronouncing dictionary for each word in the transcript, and substituting the phonetic transcription for the word (arbitrarily taking the first pronunciation if several are listed). The acoustic data for each training sentence is then divided equally among all the phones in the transcription for that sentence, and means and covariances and transition-parameters estimated from these first (almost certainly very errorful) samples. Repeated reestimation and resegmentation leads to a better and better alignment of transcriptions and acoustic data, and after a number of reestimations, room is given for amending transcriptions of particular words that have several alternative pronunciations in the dictionary, that alternative being chosen which gives the greatest likelihood in the light of the models trained thus far. Training which proceeds from an orthographic base in this way is referred to as *flat start* training.

The reason flat-start training is possible is that just as a phone may be represented by a HMM with several states, a sentence in turn may be considered as a HMM with states consisting of all the phone-models that make it up. And just as the most probable path may be found through the states of a HMM representing a phone, the most probable path may be found through a concatenated series of HMMs representing the phones of an entire utterance.

There is no reason in principle why flat-start training could not be used in FURIDA, but it remains to be demonstrated how we could ensure that the segmentation arrived at would be based on the kinds of phonetic entities FURIDA is designed to identify. Clearly, phone-boundaries no less than subphone-boundaries must be subject to alteration if flat-start training is to be possible, and the current algorithm for subsegmentation of vowels relies on an accurate placement of phone-boundaries; whether phone-level models could be trained in the first instance to yield such boundaries is not certain. There would also be problems to solve in connection with variable realisations of phones such as stops, once reliance on a phonetic transcription was removed. Use could no doubt be made of probabilities of phonetic realisation, in deriving the initial phonetic transcription from the textual one; as an example, a stop preceding a second stop (as in a transcription of "at times") will in the great majority of cases be realised by a closure subphone only, and if such probabilities were available from statistical surveys of English pronunciation, they could be used to derive a good initial transcription in subphonic terms (in subphonic terms, at least, where this was advantageous).

### 4.7.3 Duration Modelling in HMM and in Hidden Semi-Markov Modelling

It is a universally acknowledged weakness of HMM that the Markovian assumption is quite unnatural for speech, though it is also widely held that the weakness is not a cause for major concern. Having learnt in training that the self-transition-probability for some state $s_i$ in some HMM, a probability which we represent here as $a_{ii}$, is some $n$, the probability of remaining in state $s_i$ for $N$ frames is then given by

$$p_i(N) = a_{ii}^{N-1} \times (1 - a_{ii}), \mathrm{N} = 1, 2, \ldots. \tag{4.16}$$

Thus with a learnt transition-probability of $a_{ii} = 0.4$, we would have, for example,

$$p_i(2) = 0.4 \times 0.6 = 0.24,$$

$$p_i(3) = 0.4^2 \times 0.6 = 0.096,$$

and

$$p_i(6) = 0.4^5 \times 0.6 = 0.006144.$$

The duration-model for the state is quite inappropriate for speech (where the most probable duration is certainly not usually the shortest possible one, and where the probability of continuing in a given state would not decrease in this way). Given that scoring involves multiplication of transition-probabilities and spectral probabilities at every frame (or the equivalent of this in the form of additions in the log domain), one is tempted to think that the accuracy of phonetic classification using HMM should be capable of being bettered.

The reason the weakness of the duration modelling is widely maintained not to be serious is that the effect of transition probabilities tends to be overwhelmed by that of the emission probabilities in scoring: the emission probabilities tend to be very much smaller numbers than the transition-probabilities, and in the (negative) log domain this means that we are generally adding relatively small numbers (for the transition parameters) to relatively large ones. Hence it is argued that the poor duration model implicit in HMM is of little consequence – final scores are dominated by the emission probabilities (Young 1996).

This is in some ways a rather curious defence of the duration-modelling of HMM. The defence is that the admittedly poor model doesn't do too much harm, but one would think that the main reason for attempting duration modelling in the first place is that it may bring some additional good! Moreover, there remains a suspicion about the case of close competitors in the spectral domain – even small differences in duration probabilities may turn out to be significant when there is relatively little to choose between competitors in respect of their emission probabilities.

One way of trying to better the performance of HMM is by dispensing with the Markov assumption in respect of the underlying state sequence, taking the state-sequence to be a semi-Markov process, in the following sense: transitions between non-identical states will still be governed by learnt transition-probabilities, but

self-transitions — amounting in effect to periods of occupancy of a particular state — are governed by probability density functions learnt from the training-data (Russell & Moore 1985; Russell & Cook 1987). Rather than probabilities of state-occupancy decaying exponentially from a peak probability for a single frame, as in HMM, this allows for greater probabilities for durations closer to the means established in training. This modification of HMM, known as Hidden Semi-Markov Modelling (HSMM), affects only one of the two processes involved in HMM, namely the underlying state sequence. HSMM involves a significant increase in complexity compared with HMM, however ((Huang *et al.* 1991)).

The most convincing attested examples of which I am aware of superiority of HSMM over HMM (the two papers cited above) are for discrimination of minimal pairs where in practice durational differences are either the sole distinguishing cue or at least the most important, as in *pod, pot; league, leek; close (v), close (adj/n); five, fife; hard, heart; heard, hurt; killed, kilt; rider, writer; robe, rope; wand, want.* Given that the main difference between pair-elements in most of these cases will be in the duration of a vowel steady-state, HSMM would — assuming an alignment between that state and some state of the HSMM — learn two different mean durations for the two elements of each pair, and scoring of unlabelled speech could be expected to reflect the respective fits and give good discrimination, the *spectral* differences between the vowel steady-state in each pair being fairly minor. In HMMs (as opposed to HSMMs), again assuming an alignment of the kind just assumed for the HSMM, discrimination is likely not to be so good: training should result (given our assumption) in the self-transition probability for this state being a little higher for the pre-[+VOICE] than for the pre-[-VOICE] vowel, but in recognition discrimination is likely to suffer because of the neutralising effect of the geometric state-duration pdf in the pre[+VOICE] vowel with its longer duration, which would have the effect of bringing its score closer to that of the pre-[-VOICE] vowel.[17]

It should be noted, though, that the duration-model implemented in HSMM is itself far from ideal. There is a basic problem with using the durations of

[17]The thoughts expressed in this paragraph are my own, and are not in any way attributable to the authors of the papers cited in the preceding paragraph.

phones in training-data to create duration-models to be employed as aids to phonetic classification, and it is this: speech not being a bare phonetic process, but rather a *linguistic* one, the manner in which phones are produced is affected radically by things other than the simple need to give them their distinctive sound-qualities: quite apart from the need to give polysyllabic words their appropriate stress-patterns, phone-production is affected by durational changes associated with syntactic phrasing, informational load, and the expression of feeling or attitude, by the degree of hesitancy or confidence on the part of the person speaking, and no doubt by other things besides. These things become particularly evident when listening to spontaneous speech, where the ease with which the speaker articulates his or her message can vary quite dramatically from phrase to phrase and from sentence to sentence. While it would of course be going too far to dismiss out of hand the significance of measures of mean duration and dispersion for different phonetic classes, it would be even more foolish to forget that the other, less strictly phonetic features of spoken language may and frequently do have durational consequences which overwhelm the narrowly phonetic durational features of different speech-sounds, as represented by mean values, and that — for spontaneous speech in particular — statistical measures purporting to be measures of the latter are likely to be distorted by factors associated with the former. (It is perhaps significant that in the two papers referred to, the superiority of HSMM over HMM was demonstrated for isolated word recognition, where none of the sources of variability just considered come into play.)

Probably the most dramatic effects on durations of phones arise from pre-pausal lengthening – where the tempo relaxes as a major clause- or sentence-boundary is approached, and from "hesitation-hold" – where a speaker lingers on a particular sound while deliberating how to continue. Perhaps if special measures were adopted for these cases, the remaining variabilities could be handled tolerably if imperfectly well using phone duration models of the kind employed in HSMM, but it seems probable that higher-level structures such as syllables need to be invoked to get closer to really useful duration-modelling. HSMM has not been widely adopted, in any case, perhaps in part because of its increased computational complexity, and perhaps because HMM's weaknesses in regard to

duration-modelling have been masked or rendered less significant by the integration of phonetic classification into the larger task of textual transcription, where, for example, word-bigram, word-trigram, and even word-quadrigram grammars are used to constrain the transcription process: one may have less cause to worry about whether one has a single [s] or a pair in a given stretch of the acoustic record when one has such constraints pointing to a transcription as (say) "This seems to be correct".

## 4.8 Incorporating Duration-Modelling in FURIDA

In the basic implementation of FURIDA, no explicit duration-modelling of any kind is employed except in the case of stop-closures (whose circumstances are somewhat peculiar), but it seems very probable that intelligent duration modelling would enhance the system's performance significantly. The difficulty is in knowing how to achieve modelling of this kind. One problem concerns the point mentioned above, that phone-durations need to be considered in a wider context than that of mere phonetic identity. Another problem concerns the best way to combine spectral and durational scores.

### 4.8.1 Where Does the Lack of Duration Modelling Tell?

Among cases where it is either obvious or probable that the absence of duration-modelling from FURIDA is responsible for errors, we may distinguish two broad categories. In the first category, the errors are of substitution (misclassifications of one phone as another, the segmentation being more or less correct), where duration may be argued to be the sole or at the very least one of the most important cues to discrimination. Examples of this are substitutions of [aa] for [uh] — where the cepstral representation leaves little or nothing to choose between the candidates if duration is ignored — and substitutions of [-VOICE] for [+VOICE] members of fricative-pairs ([s] and [z], etc.) before [-VOICE] sounds, where the duration of the vowel or sonorant preceding the [+VOICE] fricative is probably the principal cue to the VOICE of the latter, the fricative itself typically being largely or wholly devoiced. In the second broad category of cases, the errors are

chiefly of insertion (the positing of a phone not actually present), arising from local similarity between the acoustic record at that point and a pattern reminiscent of the usual acoustic realisation of some other phone, this reminiscence often being close only if one abstracts from duration. Examples are insertions of [ii] in the velar pinch of onsets of vowels like [e] following a velar.

Turning first to the first category of cases, and taking the first example first, namely confusions between [aa] (as in "carp") and [uh] (as in "cup"): these clearly do point to the relevance of duration, if the vowels are to be discriminated at the level of phonetic classification. However, while duration-modelling based on simple collection of data without annotation for higher-level conditions might *reduce* the number of substitutions, it would no doubt yield incorrect results in particular cases (as with [uh] occurring as the last vowel before a major clause-boundary, or one sustained while the speaker hesitates as to how to continue, and with an [aa] occurring as part of a hackneyed, rapidly delivered phrase).

The second example given for the first category of cases was the substitution of (e.g.) [f] for [v] before a [-VOICE] sound. One thing that looks as if it might work in our favour here is that before a [-VOICE] phone, a [+VOICE] fricative *has to* belong to the same syllable as the vowel or sonorant that precedes it (assuming that we are not dealing with a disfluency of some kind). While the general point has been insisted on above that the duration of any phone needs to be looked at in the light of the linguistic context of its occurrence, when we are given at least some insight into the suprasegmental context as here there would appear to be scope for relativistic (phone-to-phone) duration-modelling of a kind that would appear to be helpful: there will be a proportionality between the duration of the fricative and that of the preceding vowel or [dl] or nasal, which may be expected to differ from the corresponding proportionality when the fricative is a [-VOICE] one, regardless of speaking-rate or sentence-internal position (both of which should affect both phones equally). Unfortunately, the simple comparison between relative durations in this case and similar durations in the case where the fricative is [-VOICE] is not readily available, because the possible syllabic affiliations in the latter case are several, and different prosodic and syntactic contexts will therefore result in a variety of relative durations. Whereas a [-VOICE] fricative belonging to the same syllable as the preceding vowel or sonorant will show a

different proportionality than its [+VOICE] counterpart in the same context, one which begins a new syllable or word (such as an [s] before a [t], [p] or [k]) might not. The situation can be seen to be even more difficult when one considers that we may have [z s], [v f], [zh sh] and [dh th] sequences where there may be no real acoustic counterpart (within the fricative stretch) of the phonological distinction, raising the possibility of errors of insertion or deletion (failure to recognise the presence of a distinct phone) which further threaten the possibility of getting reliable relative durations. Facts such as these would seem to point to the need for a syllabic or similar network within which to do DP scoring – a network within which all possible affiliations could be explored in parallel.

I turn now to the second category of errors, where there may be room for doubt as to whether duration-modelling is the most likely answer. Consider again the insertions of [ii] at the beginning of [e] vowels following velars. True, the inserted [ii]'s are typically very short, so that even naive duration-modelling (insensitive to utterance-level features) might well block the insertion. On the other hand, such modelling would probably introduce errors of its own elsewhere, since extremely short realisations of phones — including vowels like [ii] — do occur. It is in any case arguable that the root of the insertion-problem here is not the lack of duration-modelling but rather the modelling by means of a single gaussian: in the population of vectors assigned to the [g_eB] subphone in training, vectors with acute velar pinch (so to speak) will constitute a minority of perhaps only 10 per cent or so of the total, and such vectors will therefore not get the probability they deserve in recognition, given the normal model. With a bimodal distribution, the problem of [ii]-insertion might conceivably be solved without the need for duration-modelling, though it must be conceded that it is perhaps equally possible that with such a bimodal model some *true* [g ii e] sequences might get classified as [g e] sequences, if no duration-modelling was in use. An alternative solution might be to model the onset of post-velar [e] — and similarly for all analogous cases — with two subphones, each using a single gaussian for spectral modelling.

Some form of relative *within-phone* duration-modelling could be introduced either as an alternative to or, perhaps more promisingly, in addition to one or other of these proposals. If two subphones — [g_e1B] and [g_e2B], let us say

— were used to model the onset of post-velar [e], then in the genuine cases the relative duration of the two subphones might be expected to tend toward a mean of perhaps 1:4 or 1:5. Assuming that all pairs of consecutive within-phone subphones were being scored in the same way with respect to their relative durations (vis-a-vis one another), misclassification of genuine [g ii e] as [g e] (i.e. as [g_e1B g_e2B Ce ...]) would probably be prevented: in a true [g ii e] sequence, any spectral similarity between the [ii] vowel and a [g_e1B] subphone might be counterbalanced by a duration-based cost because of its long duration relative to the [g_e2B] subphone that would also have to be hypothesised. (The insertion of [ii] in a genuine [g e] case, meanwhile, we would hope to prevent by the two-subphone modelling of post-velar [e] onset, rather than by the presence of the relative duration-modelling; the latter would presumably do little of itself to block the insertion, since a single frame [g_iiB] followed by a single frame [ii_eA] might well amount to a fairly probable relative duration (50:50) for these two subphones.)

Given reliable subphone-to-subphone relative duration statistics, the mechanics of incorporating score modification based on them into the DP stage would not appear to present insuperable problems, certainly if attention is restricted to pairs of consecutive subphones within phones; it would be necessary only to record the number of frames on each exit from a subphone, for consultation at each candidate for final frame of the next subphone in the path. On the other hand, obtaining such statistics in the first place would present some problems, given that samples will be counted in numbers of subphones rather than in numbers of frames as was the case for spectral modelling. It might well turn out to be the case, though, that relative (within-phone) subphone-durations would prove to be fairly uniform for speakers of a given accent, and perhaps even across accents in most cases, so that with the increasing availability of labelled speech databases, the requisite statistics could be gathered fairly readily.

The general validity of some of the ideas expressed here about relative subphone-durations is not, of course, guaranteed. Lengthened pre-pausal vowels following [+VOICE] sounds appear to show lengthening in all three subphonic stages (but see (Beckman & J.Edwards 1992) for evidence that most lengthening occurs in the core and offset). Whether it is correct to claim that "stretching" and "squashing"

of phones in other positions affects all their subphonic stages roughly equally is at least not certain. The research literature as summarised by Lindblom (Lindblom 1983) suggests that there are alternative ways of producing vowels more quickly, for example: either one can reduce the extent to which one attains the vowel's target-value ("undershoot"), or one can move toward and away from the target more quickly; Lindblom quotes Kuehn and Moll (Kuehn & Moll 1976) as saying that at a rapid speaking-rate, "speakers have the option of either increasing velocity of movement or decreasing articulatory displacement". Where undershoot occurs, we can expect the pairs of subphones subjected to relative duration scoring to be onset and offset subphones, and it seems not unreasonable to think that the proportionalities for such pairs might be similar whatever the speaking-rate or other material conditions. But where the option is taken to increase velocity of movement to a target, would the same sort of proportionality be maintained between onset and core, or between core and offset, as would be found for "normal" durations? Presumably in fast speech with rapid onset and offset, but with attainment of the target-value, one would still expect the core or steady-state to be maintained for a relatively brief time, in which case the proportionalities would indeed appear to have some chance of being preserved, but this is clearly a case where one would have to go and take measurements.

### 4.8.2 Some Questions Regarding The Mechanics of Duration Scoring

I turn aside now from issues involved in attempting duration-modelling of a kind sensitive to all the linguistic factors affecting duration, and consider a number of questions pertinent to the "nuts and bolts" of duration-scoring, and in particular to the integration of duration- and spectral-scores.

Suppose that subphonic duration-distributions had been estimated from large amounts of training-data (based on counting the numbers of frames in each subphonic token in the final alignment of transcription and training data at the end of the final iteration in training). The estimates could be represented as histograms, or probabilities could be pre-computed for all possible frame-counts

from a parametric distribution of some kind; in either case, applying duration-costs in DP search might involve nothing more costly than a look-up for the given class and frame-count, and adding in the negative log probability to the sub-path score. But a number of questions hang over this suggestion. One question which surfaces immediately is whether duration-costs should be applied only on leaving a subphonic sub-path, or whether each "self-transition" – each additional frame spent within a sub-path through a particular subphone – should also incur some cost, as in HMM. At any class $i$ at frame $j$, when looking for the best predecessor $k$ at $j-1$, if we stipulated that $k$ should incur additional cost only if it is distinct from $i$, could we be in danger of biasing the decision at that point in favour of self-transitions? If "self-transitions" too are to incur costs, what should those costs be?

If it were decided that self-transitions should be penalised along with exit-transitions, it would be simple enough to do this in a way that appeals to reason. Suppose we are working with look-up tables that give for each class the probability of a class-token being precisely $n$ frames long (where $n$ ranges from 1 to, say, 30). In the course of scoring, we could impose costs on self-transitions as follows: in extending a path from $k$ at $j-1$ to $i=k$ at $j$, where the frame-count at $j-1$ is $n$, we ask what probability there is that a token of class $k$ could be more than $n$ frames long, and add the negative log of this probability to the path that extends to $i=k$ at $j$; the probability of a duration of more than $n$ frames is of course equal to 1 minus the probability of a duration of $n$ or fewer frames, so that once again the requisite costs could be precomputed (from the subphone-specific cumulative distribution functions) and stored for each class for look-up. Clearly, the costs of self-transitions would then bear some natural relation to actual durations of subphones in the training-data (as in HSMM, and in contrast to HMM), though whether "self-transitions" should bear any cost at all must depend on the overall scheme for duration-scoring.

What then of exit-transitions (the costs to be incurred when the candidate best predecessor $k$ for class $i$ is distinct from $i$)? Should such a transition incur costs proportional to the probability of a token of class $k$ being precisely $n$ frames long? Or, if self-transitions are penalised in the way suggested in the previous paragraph, should the total penalty for the self-transitions that took place prior

to the exit be deducted from the exit-penalty for the given $n$? Is there any chance that the typically fairly small penalties for duration will be powerful enough to give durational considerations the voice we might think they deserve to have, and so be able in some cases to override cepstral probabilities? Should duration costs be weighted in some way to make them more influential (at least if the other problems involved in duration-modelling could be solved)? The uncertainty as to whether some such weighting should be applied is connected with uncertainty regarding the precise connection between – on the one hand – *frame-level* spectral or cepstral probabilities, and – on the other – duration-probabilities that apply (in the case of "exit" transitions at least) to *sequences* of frames.

## 4.9 Conclusion and Summary

I have described the techniques used in FURIDA for training sub-phonic models and for carrying out phonetic transcription using these models. I have attempted to clarify the relationship between the techniques described here and those of Hidden Markov Modelling. The primary motivation for undertaking the work that led to FURIDA was to develop a technique that made clear and explicit connections with phonetic phenomena, and it is my belief that these techniques open up another front, so to speak, in the battle for improved recognition, a front which may prove more attractive to the phonetically minded than that dominated by the rather black box of HMM.

I have pointed to the well-known weakness of the duration-model in HMM, and also to the limitations of the duration-modelling found in HSMM, which is insufficiently sensitive to linguistic factors other than mere phonemic identity. However, I have also acknowledged the difficulties in the way of incorporating intelligent duration-modelling in FURIDA, which currently has no duration-model to speak of. I have discussed in a tentative way the possibility of using relative duration-modelling to resolve a number of problems where duration appears to be important; in some cases it is relative durations of *phones* (the proportional lengths of a vowel and a following fricative, for example) that are of interest, and in others the relative durations of *subphones* (the proportional lengths of the subphones of a single vowel, for example). Such modelling would seem to require the

integration of phonetic classification into a framework built around higher-level structures such as syllables; the accommodation of pre-pausal lengthening would require recognition of entities such as clause and sentence; and pauses, hesitation, "hesitation-hold", and the like would also have to be modelled. I incline to the belief that such accommodations would be possible without going beyond the syllabic to the lexical level (a clause, for example, being modelled simply as an entity which may have the effect of slowing the tempo at which sounds are articulated as it comes to an end), but this has not been demonstrated here.

Finally, I raised a number of questions concerning the mechanics of basic duration-scoring. These questions point to the need for an integrated theoretical framework relating spectral frame-level scores on one hand and durational subphone-level scores on the other. To date, I am able to offer no worthwhile contribution to this issue.

# Chapter 5

# Making the Most of Limited Training Data

## 5.1 Introduction

The main focus of this chapter is on the strategies available for combining data from different classes when there is insufficient data to work with specific statistical parameters for each class individually. The approach used in FURIDA is first described in some detail. Two clustering algorithms that have been popular in ASR are then described. Finally, some further options for dealing with data-limitations are considered briefly.

For any statistical classifier to work well, it goes without saying that the data-samples from which the class parameters are estimated should be representative of their classes. For the particular application to speech recognition, meanwhile, we need to use context-conditioned classes if we are to get good results (this is particularly true when working with single gaussians as here). There is thus a conflict between two basic requirements, in that the more we distinguish according to context, the more we divide the available data into distinct classes, so that the average amount of data for each class decreases, thus increasing the risk that some classes will be poorly represented by their samples. Some form of compromise has therefore to be sought. Roughly, the compromise typically reached involves identifying (treating as one) any set of contexts $C$ that have similar effects on the

acoustic realisation of phones of a particular phonetic class [$\alpha$], and so clustering species of [$\alpha$] that are annotated as falling in any one of the contexts included in $C$. Ideally, one would like to find the *optimal* compromise — the best possible (least generalised) set of classes given the available training data, and a number of strategies have been developed in ASR for achieving just this (5.4).

While the primary purpose of clustering (merging, generalising) is to create classes that have a good chance of being accurately represented by the training data, clustering may still be an appealing move even when one has sufficient data to model a number of classes specifically, if there are advantages to be gained in computation-time, memory requirements, or even in reduction of the error-rate, by treating them as a single class. Consider, for example, the subphones [s_iiA] and [s_iiD1A];[1] in the absence of any evident acoustic difference between these two classes, it might be decided (as it was in FURIDA) to merge them as a matter of course (in FURIDA, to form [s_IIA]), without even considering the question of whether the amounts of data made this necessary or not. Two issues are cued by this that will remain in the background, so to speak, throughout most of this chapter. One concerns the question of subjective vs objective criteria for deciding whether there is an acoustic difference that matters between a pair of subphones, and for deciding on the relative similarity of different pairings of subphones. The second concerns the influence on decisions regarding clustering of the *approximate* amount of data available; with only 200 training-sentences available, one will be more inclined to look for immediate merges (as of [s_iiA] and [s_iiD1A]) than one might if 50,000 training-sentences were available. In section 5.3 I describe in some detail the generalisation strategy employed in FURIDA, while in section 5.4 I consider clustering strategies that have been used in state-of-the-art recognition-systems that normally operate with very large amounts of training-data. In FURIDA, with between 190 and 250 training-sentences, a great deal of generalisation was executed *ab initio*, or very nearly so, because it was obvious that it would be necessary. Moreover, judgements about what merges were reasonable — whether the merges were to be executed as a matter of course, or conditionally upon data-shortages making them necessary — were made on

[1]The first of these is the offset of an [s] preceding an [ii], the second the offset of an an [s] preceding the first element of an [i@] diphthong, as in "this ear".

the basis of familiarity with the spectrographic record, with some extrapolation on the basis of phonetic knowledge or phonetic plausibility. With hindsight, it now seems to me to have been a mistake to attempt recognition using such a limited training-set. With a significantly larger set of training-data, it would make sense to use an automatic or quasi-automatic approach to clustering, such as those described in section 5.4. The detailed exposition in section 5.3 is therefore given more because of the relevance of the clustering procedure I employed to the evaluation of FURIDA's performance (as detailed in the next chapter), than because of any conviction that the procedure represents a significant contribution to the field of ASR.

## 5.2 Minimum Requirements for Statistical Modelling

As described in 4.3, class-specific covariance matrices are used to calculate discriminant scores, the covariance matrices being estimated using the Maximum Likelihood formula

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t, \tag{5.1}$$

where $n$ is the number of vectors available for the class. In order to be able to compute discriminant scores we have to be able to invert $\hat{\Sigma}$, from which it follows that for vectors of dimensionality $d$ we need at least $d+1$ vectors to proceed (if the number of vectors for the class is smaller than this, $\hat{\Sigma}$ will be singular and so will not have an inverse, making it impossible to proceed (Duda & Hart 1973)).

Given a 16-element representation vector, we thus have an absolute minimum requirement of 17 vectors for any class for which we are to construct a statistical profile. As Duda and Hart comment, however, "To smooth out statistical fluctuations and obtain a really good estimate, it would not be surprising if several times that number $[d+1]$ ... were needed" (p.68). In fact, if one reflects for a moment, one readily sees that for the particular application in view, several times the number of $d+1$ are almost *bound* to be needed, and would certainly at the very least be highly desirable. This is because we are dealing with mere

slivers of the acoustic record of speech, so that a class sample of even 30 vectors could be provided by just three or four occurrences of the relevant phonetic event (for example by just three or four offsets of [aa]-tokens preceding [sh]-tokens); given all the ways in which such events can vary — in inter-articulator timing, in signal strength, in speaking-rate, and so on — it would probably be better if we thought not in terms of several times $d+1$ vectors, but of several times $d+1$ tokens of the phonetic event in question!

For present purposes, acting in such a sensible manner was not feasible, and a figure of 30 was generally taken as the minimum number of vectors for statistical profiling.[2] In the rest of this discussion, the term *quorum* will be used to mean "the minimum number of class-vectors for statistical modelling" (though I shall also sometimes use it more loosely to mean a collection of class-vectors equal to or greater than this in number). Once a quorum is set, the question arises of what to do about the large number of classes with sample-sizes that fall below it.

## 5.3 Generalisation of Classes in FURIDA

### 5.3.1 Basic Principles of Generalisation

Two different forms of generalisation are used in FURIDA (as already hinted at in the introduction), namely immediate and conditional. Immediate generalisation (generalisation which takes place as a matter of course, regardless of sample-sizes) may be implicit in the the basic class-inventory; even the use of left-sensitive onset subphones and right-sensitive offset subphones may be seen as a form of immediate generalisation, and as will be seen in 5.4, entities very close in kind to these are in fact arrived at in leading ASR systems as a *result* of automatic clustering procedures applied to states of HMMs. Some generalisations which I am calling immediate are not implicit in this sense, but are immediate in the sense that they do do not wait upon data-shortages, and are executed as a matter

[2]The first time I got FURIDA to work at all, I did so with this figure set at only 18, just 2 greater than d, and was plunged into despondency by the results; after doing nothing other than increasing the figure to 24, and thereafter to 30, I was relieved to find a *dramatic* improvement, even though the higher figure forced generalisation to a higher degree than I had first thought desirable.

| SPECIFIC FORMS | IMMEDIATE MERGE |
|---|---|
| [$\alpha$_aB], [$\alpha$_aD1B] | [$\alpha$_AB] |
| [$\alpha$_eB], [$\alpha$_eD1B] | [$\alpha$_EB] |
| [$\alpha$_aaB], [$\alpha$_aaD1B] | [$\alpha$_AAB] |
| [$\alpha$_ooB], [$\alpha$_ooD1B] | [$\alpha$_OOB] |
| [$\alpha$_iiB], [$\alpha$_iiD1B] | [$\alpha$_IIB] |
| [ii_$\alpha$A], [iiD3_$\alpha$A] | [II_$\alpha$A] |
| [i_$\alpha$A], [iD3_$\alpha$A] | [I_$\alpha$A] |
| [ax_$\alpha$A], [axD3_$\alpha$A] | [AX_$\alpha$A] |
| [e_$\alpha$A], [eD3_$\alpha$A] | [E_$\alpha$A] |

**Table 5.1.** Immediate Merges of Vowel Subphones

of course; further details will be given shortly. Conditional generalisations, on the other hand, take place only as and when data-shortages make them necessary.

The most notable cases of immediate merges for vowels are listed in table 5.1 (where $\alpha$ represents any consonant).

With consonants, too, a number of generalisations take place automatically, regardless of sample-sizes. Most of these have an obvious rationale. For example, right subphones of a consonant $\alpha$ that have a [b*c] as right context (where [b*c] may be any of [bc], [bbc], [bdc], [bgc], [btc], etc.) are merged automatically to form [$\alpha$_bcA], while left subphones of a consonant $\alpha$ that has a [*bc] as left context are automatically merged to form [bc_$\alpha$B]. Similarly, left subphones of a consonant $\alpha$ with either [sh] or [chb] as left context are automatically merged to form [SCH_$\alpha$B], and left subphones of a consonant $\alpha$ with either [zh] or [jhb] as left context are automatically merged to form [ZJH_$\alpha$B], while right subphones of a consonant $\alpha$ with [l], [cl] or [lo] as right context are automatically converted to [$\alpha$_NDLATA].

For conditional generalisation, the most obvious procedure to adopt would appear to be to merge subphones which are affected in similar ways by their context. For example, offsets of [s] before [tc], [dc] and [n] are all very similar because of the common PLACE feature, so that if any of [s_tcA], [s_dcA] or [s_nA] prove to have less than a quorum of vectors, it makes sense to merge first of all with one or both of the others to form a generalised class, say [s_ALVA] (offset

of [s] before an alveolar consonant), and to estimate parameters for the generalised class (assuming it itself has a quorum of vectors, failing which, further generalisations will have to be sought). With vowels, however, the scope for generalisation purely according to consonantal context proves to be limited because of the highly specific effects associated with different contexts; PLACE effects on vowel-offsets, for example, are cut across by variations arising from differences in glottal state (breathiness, glottalisation-effects) associated with different values for VOICE and MANNER, and by variations resulting from the presence or absence of nasality. What then is to be done when data-supplies will not allow specific modelling? One possibility is to merge onsets or offsets of *different* vowels that have a *constant* consonantal left or right context, according to the degree of closeness of the vowels in vowel space (e.g. to merge [aa_dlA], [uh_dlA] and [o_dlA]). In an earlier version of FURIDA this approach was tried for all vocalic subphones in consonantal contexts, but found to yield poor results because of the frequent need for very extensive generalisation, often leading to subphones such as [BACK_$\alpha$A], [$\alpha$_NON-BACKB], or even [VOWEL_$\alpha$A], [$\alpha$_VOWELB] — hardly sensible candidates for modelling with single gaussians! Ideally, perhaps (for a system operating with a very meagre supply of training-data), an automatic procedure would be used which had the option of considering both forms of merge — generalisation according to the consonantal context, or generalisation on the vowel itself according to closeness in the vowel-space, with the consonantal context constant — at each step of the generalisation-process, using an objective criterion to decide which form to adopt at each step. In the present case it was not deemed a wise use of time to try to develop such an algorithm, and the following short-cut was used for generalising vowels in consonantal contexts: vowel onsets and offsets with consonantal contexts are merged immediately according to consonantal context wherever there is scope for doing so, so that for example, [o_ngA], [o_gcA], [o_gbA], [o_kcA] and [o_kbA] are merged to form [o_VELA] (offset of [o] before a velar). If a quorum is still not available after such a merge, further conditional generalisation takes place by merging vowel onsets or offsets with the given consonantal context, so that, to continue with the example, [o_VELA] would be merged first with [uh_VELA] and [aa_VELA], and if a quorum was still not reached, the merged class would be further merged

with offsets of other back vowels with velar right context.

These immediate merges for vowels (effected with a view to limiting the need for subsequent conditional merging of different vowels) are given in table 5.2 and 5.3 (the reader in a hurry may be content to note that immediate generalisation is basically according to the PLACE of the consonantal context — with variable attention to distinctions of VOICE as between onset and offset subphones — but is urged to read the more detailed comments that follow).

Some explanation is called for for the obvious asymmetries in table 5.2 and 5.3 (in the generalisations for onsets, VOICE has been ignored, while in the case of offsets it has been taken account of everywhere except in the case of velars). The combining of [-VOICE] alveolar and denti-alveolar contexts also merits some comment.

The first point to note is that if there were a hundred or more representation-vectors for each specific case, few if any of these immediate merges would be irresistible; for the most part they represent decisions taken under duress (because of lack of data). For example, returning to a point touched on earlier, it is certainly not ideal (particularly given the simple modelling scheme (single gaussian)) to merge onsets or offsets that have nasal contexts with others that do not, given the more or less unique effects on vowels of late or early velum-movement in nasal contexts.

Given that some merging according to consonantal context is inevitable, then, it is a fact (for English at least) that VOICE is in general far more significant for offsets of vowels than for onsets, *if we think of onsets and offsets as involving discernible formant-movements.* In the case of onsets, the most dramatic effect a [-VOICE] left context can have is to ensure that no onset appears at all, and the system is, as has been described, designed to allow for immediate transition from a [-VOICE] phone to a vowel core; thus the *actual* onsets of a vowel $\alpha$ that has a [-VOICE] left context, as represented by actual subphonic classes, will be just those that do show a measure of formant-movement (those with relatively early onset of voice, relative to articulatory repositioning for the vowel), and so have something in common with onsets of $\alpha$ following [+VOICE] left contexts with the same or similar PLACE feature. With vowel-offsets, on the other hand, while the *direction* of any formant-movement will tend to be similar for [+VOICE] and

| SPECIFIC FORMS | IMMEDIATE MERGE |
|---|---|
| [zh_αB], [jhb_αB], [y_αB], [sh_αB], [chb_αB] | [PAL_αB] |
| [bb_αB], [*bc_αB], [bnc_αB], [pb_αB], [*pc_αB], [v_αB], [f_αB], [m_αB], [msyl_αB], [mo_αB], [Pb_αB] | [LAB_αB] |
| [gb_αB], [*gc_αB], [gnc_αB], [kb_αB], [*kc_αB], [ng_αB], [Kb_αB] | [VEL_αB] |
| [db_αB], [*dc_αB], [dnc_αB], [tflap_αB], [tb_αB], [*tc_αB] [Tb_αB], [n_αB], [nsyl_αB], [no_αB], [s_αB], [z_αB], [th_αB], [thR_αB], [dh_αB] | [ADALV_αB] (alveolar or denti-alveolar left context) |
| [l_αB], [cl_αB], [lo_αB], [plb_αB], [klb_αB], [tlb_αB], [blb_αB], [glb_αB], [dlb_αB] | [NDLAT_αB] |
| [r_αB], [trb_αB], [Trb_αB], [prb_αB], [Prb_αB], [krb_αB], [Krb_αB], [drb_αB], [brb_αB], [grb_αB] | [R_αB] |
| [α_bbA], [α_bncA], [α_b*cA], [α_blbA], [α_brbA], [α_vA], [α_mA], [α_msylA], [α_moA] | [α_VLABA] |
| [α_gbA], [α_gncA], [α_g*cA], [α_glbA], [α_grbA], [α_ngA], [α_ngsylA], [α_ngoA], [α_kbA], [α_k*cA], [α_klbA], [α_krbA] | [α_VELA] |
| [α_dbA], [α_dncA], [α_d*cA], [α_dlbA], [α_drbA], [α_drcA], [α_tflapA], [α_zA], [α_nA], [α_nsylA], [α_noA], [α_zhA], [α_jhbA], [α_jhcA] | [α_VCORA] ('COR' stands for 'coronal' (used rather loosely here)) |
| [α_pbA], [α_p*cA], [α_plbA] [α_prbA], [α_fA] | [α_NVLABA] |

**Table 5.2.** Immediate Merges of Consonantal Contexts of Vowels

| SPECIFIC FORMS | IMMEDIATE MERGE |
|---|---|
| [α_tbA], [α_t*cA], [α_trcA], [α_tlbA], [α_sA], [α_chcA], [α_chbA], [α_shA], [α_thA] | [α_NVCORA] |
| [α_lA], [α_clA] | [α_NDLATA] |
| [α_GL*cA], [α_GLsA], [α_GLfA], [α_GLshA], [α_GLthA], [α_GLchbA], [α_GLhA] | [α_GLNVCA] (any glottalised offset before a [-VOICE] consonant) |
| [α_GLzA], [α_GLvA], [α_GLzhA], [α_GLdhA], [α_GLjhbA], [α_GLlA], [α_GLclA], [α_GLmA], [α_GLnA], [α_GLyA], [α_GLdlA] | [α_GLVC1A] (any glottalised offset before a [+VOICE] consonant other than [r] or [w]) |
| [α_GLwA], [α_GLrA] | [α_GLrwA] |

**Table 5.3.** Immediate Merges of Consonantal Contexts of Vowels (continued)

[-VOICE] right contexts of similar PLACE, other features such as power and breathiness will tend to be different, in certain cases even markedly different, so that not automatically ignoring VOICE makes very good sense in the case of offsets. In the case of velar right contexts, however, the problem is that if we differentiate the context according to VOICE as well as PLACE, we are left only with [kc] (and possibly a token or two of [kb]) as [-VOICE] right contexts, and so will be driven to excessive generalisation according to vowel-quality in the next stage of the generalisation-process (still awaiting detailed description).

Similarly it seemed desirable to find some way to avoid excessive generalisation on the vowel for vowels following or preceding [dh], and it was for this reason that these contexts were put into a group with alveolars.

The variable sizes of the 'pools' for each immediate merge, and consequently the variable degrees of generalisation on the vowel that may be required to reach a quorum in different cases, open up a question that will be revisited in 5.3.4, namely that of introducing bias into the classifier by providing highly generalised classes for some of the original forms, and less generalised classes for others.

Once all immediate merges have been completed, conditional merges take place on the vowel where required (that is, where classes resulting from immediate merges, or original specific classes in a few instances, are still short of a quorum). The general principle governing conditional merges has already been indicated — those vocalic subphones are merged first which are closest to each other in the vowel-space — but a more detailed description is given via examples in section 5.3.3.

Modelling transitions between members of pairs of consecutive vowels — including cases where the first 'vowel' was a 'D3' element of a diphthong — was made extremely problematic because of the sparsity of the data. Some immediate generalisation was therefore effected, using the behaviour of F2 as the principal factor; thus with a vowel like [ii] or [iiD3] or [iD3] as first member and a vowel like [oo] or [o] or [aa] or [uh] or [uu] as second member of a pair, generalisation would immediately create [H2LR] (High to Low) for the offset of the first vowel, and [H2LS] for the onset of the second vowel[3]. Following immediate generalisations

[3]The use of 'R' and 'S' suffixes in place of the usual 'A' and 'B' arises from historical factors

of this type, further conditional generalisation takes place where required.

With consonants, after the initial immediate merges according to context, all further (conditional) generalisation continues to be according to context. In no case is a subphone of one consonantal class merged with a subphone of a different consonantal class. Further details and examples are given in the next two sections.

### 5.3.2 Mechanics of Conditional Generalisation

The perceptive reader may already have observed that generalisation-procedures need to be invoked at the beginning of each stage of the iterative resegmentation process during training, and again on the completion of the process, and since the nature of that process entails that data-counts for individual classes may vary from iteration to iteration, the procedures used for conditional generalisation must be capable of dealing with inputs that are unpredictable in terms of sample-sizes for each class. Not knowing a priori the class-inventory or the sample data-counts at each point where generalisation must take place, we require a generalisation-procedure that takes the data it finds to be available for specific classes (or classes resulting from immediate generalisation) and creates a set of classes that each have a quorum but preserve as much context-dependent specificity as possible.

The solution adopted was to create for each case (for example, for all possible right subphones of [s] with consonantal right contexts) a hierarchical structure that could be used to control generalisation in such a way that where data did not allow specific subphones to be modelled, then — to the extent that the data did allow — subphonic groupings whose specific forms manifest common features would be modelled in their stead. The possible groupings are arranged in such a way that this commonness of features (one might also say, suitability for modelling with a single gaussian) declines only as greater and greater levels of generalisation are forced by continuing data-limitations, but with always (ideally) at least a maximally generalised class available in the last resort to represent all possible specific forms that might be encountered in recognition-mode (for

---

of no intrinsic interest

example, at least an [ng_CONSA] class to represent all possible right subphones of [ng] with consonantal right context). (Generally "maximally generalised" means generalised up to the level of CONSONANT or VOWEL; there are no cases where the context is simply ANY (CONSONANT OR VOWEL).)

Trees are the appropriate data-structure for implementing hierarchical procedures of this kind, and the computationally convenient binary tree serves the purpose more or less adequately. (The rest of this paragraph may be skipped by a reader familiar with trees as data-structures and with methods of tree-traversal). A binary tree is a tree each of whose nodes (or vertices) has at most two nodes directly below it, with any dependent node to the left being designated as the left child of the node immediately above it, and any node to the right as the right child of the node immediately above it. Nodes with no children are referred to as leaf-nodes or terminal nodes, and all others as non-leaf or non-terminal nodes. References to parental, ancestral, descendant and sibling relationships between nodes should need no explanation. The root of the tree is ancestral to all other nodes in the tree, and any node in the tree may also be considered as the root of a subtree comprising itself and all the nodes below it (all the nodes to which it is ancestral). A path in a tree is a list of distinct nodes, with successive nodes connected by arcs of the tree. Paths will not enter into the discussion here except in helping to define the idea of *levels* in the tree. Each node of a tree lies at a particular level of the tree; the level of a node is given by the number of nodes in the path from the node to the root of the tree, excluding the node itself, as illustrated in figure 5.1.

Levels of a tree should be distinguished from levels of generalisation in the discussion that follows; firstly, lower levels of generalisation would tend to correspond with higher-numbered levels of a tree, but secondly, a leaf-node, representing a specific class or a class resulting only from immediate generalisation, may be found at any level in a tree other than the root level (this is true for the trees used in the present application, where there are no single-node trees).

Algorithms that process trees as data-structures typically involve systematic consideration of all or some of the nodes of a tree. In *tree-traversal* each node of a tree is visited exactly once, where *visiting* a node typically implies taking some action at that node, for example adding to or altering information stored at

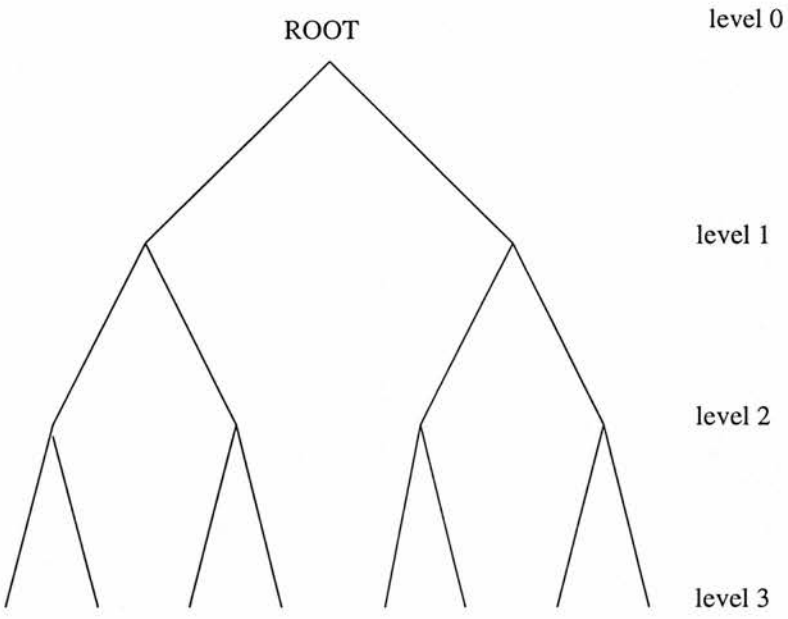



**Figure 5.1.** Levels in a Tree

the node. Assuming that we choose to follow a policy of always going left before going right, there are three possible ways of traversing a tree depth-first, and once a choice between the three is made, a precise order of visiting is fully defined. In *pre-order* traversal, we visit the node arrived at, then traverse its left subtree in pre-order, then traverse its right subtree in pre-order; in *in-order* traversal, we first traverse the node's left subtree in in-order, then visit the node, then traverse its right subtree in in-order; finally, in *post-order* traversal, we traverse the node's left subtree in post-order, then traverse its right subtree in post-order, and then visit the node.

For purposes of further exposition, I shall restrict attention in the first instance to consonant subphones (annotated for left or for right context). For each such case a (binary-branching) tree-structure is created to govern the generalisation-process. Each node in the tree is associated with a named context, the leaf-nodes being associated with specific contexts (or with contexts resulting from immediate merges), and the non-leaf nodes with contexts created as a result of conditional generalisation. The actual design of any particular tree, in the sense of the precise

placement of specific contexts with respect to each other at leaf-nodes, and the hierarchical groupings represented at non-terminal nodes, is determined as far as possible by purely phonetic considerations, and this topic will be treated in more detail via examples in section 5.3.3. In general, of course, the aim is to so arrange things that those contexts are merged first that have the most similar acoustic effects on the subphones concerned, with further merges being put back in proportion to the degree that their acoustic effects are different.

The general kind of consideration that determined the construction of trees may be illustrated by the case of tree-structures for generalising pre-consonantal offsets of (1) [+VOICE] fricatives and [jhb] and (2) [-VOICE] fricatives and [chb]. As described in Chapter 2 (2.5.1), the VOICE feature of phones following [+VOICE] fricatives and [jhb] has a very major effect on the acoustic realisation of their offsets, while the VOICE feature of phones following [-VOICE] fricatives and [chb] is of very minor importance in this respect. In the generalisation-trees for these two groups, accordingly, there is a very basic difference in design: in the case of (1) the main split from the root level (the node at root level is associated with all consonantal right contexts of whatever kind) is between [+VOICE] and [-VOICE] consonants, with PLACE and MANNER determining only subordinate hierarchies within this major dichotomy; in the case of (2), however, the principal split is used to hive off [r], [w] and laterals from everything else, and in the other main branch the first split is between stops and nasals on one hand, and fricatives, affricate-releases, [h] and silence on the other, with the division between [+VOICE] and [-VOICE] pair-elements within these groups appearing only at very peripheral levels of the tree, and much greater importance being given to PLACE.

At the commencement of generalisation, we traverse the tree (the order of traversal at this point is a matter of indifference) and at each leaf-node the number of vectors available for the subphone associated with that node (which may of course be zero) is stored at the node. Generalisation then proceeds as follows: first, traversing in post-order, numbers of vectors at each pair of child nodes are summed, and the sum recorded at the respective parent node, a process which continues up the tree so that at root level the total number of vectors for all leaf-node subphones is recorded; I shall refer to these records of data-counts

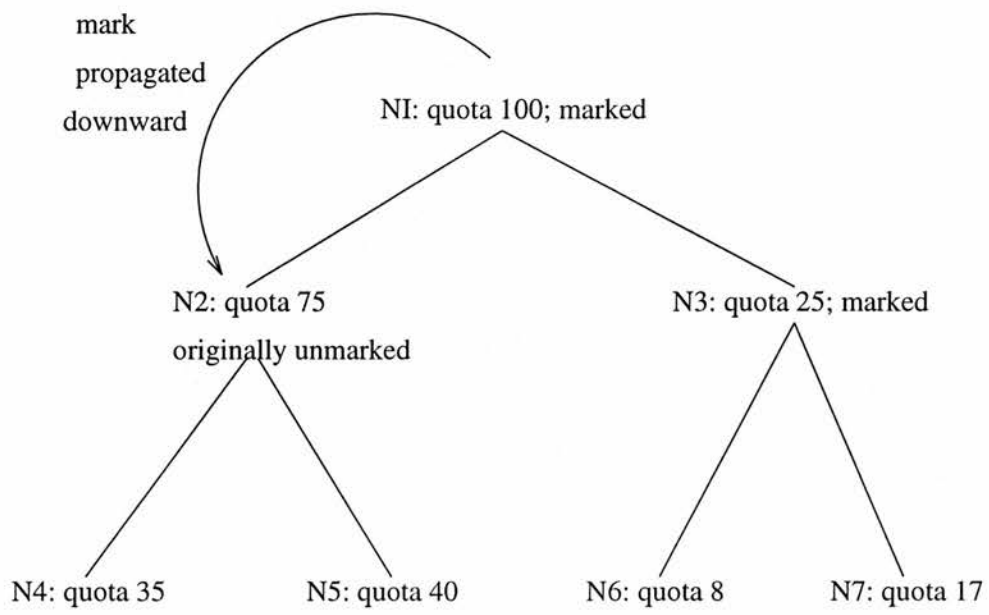



**Figure 5.2.** Downward Propagation of Marks

as *quotas*. Next, the tree is traversed again in post-order, and at each non-leaf node, if one or other of its child nodes has a quota less than the quorum, the parent-node is marked to indicate that this is the case. When this traversal is complete, the tree is traversed once again in pre-order, and the marks assigned in the previous traversal are propagated downwards from any node that is itself marked. Once this has been done, a further, post-order, traversal takes place, in which at any marked non-leaf node with a non-zero quota we combine the class-samples for the child nodes into the generalised form associated with that node. Clustering comes to a (possibly temporary) halt on visiting an unmarked node.

The downward propagation of marks takes place in the way shown in Figure 5.2 (I use a dummy tree for illustration), assuming a quorum set at 30. As illustrated here in the case of node N2, a node may be unmarked originally in virtue of neither of its children being short of a quorum, but data-deficiencies for a class associated with a node elsewhere in the tree may result in it receiving a mark when marks are propagated downward to non-leaf nodes, as happens here

with a mark originating at node N1 (a mark conferred on N1 in virtue of its child N3 having less than a quorum of vectors). A policy of "one for all and all for one" is thus being followed, with the chances of more specific modelling for some contexts (and failure to model others altogether) being rejected in favour of one which takes all classes to whatever level or levels of generality may be required to ensure that all possible specific forms have some kind of representation in the class inventory. (It may be stressed once again that the amount of data worked with here is quite unrealistic, and that with the kinds of amounts of data that may be expected in real-world ASR, generalisation could be expected to remain at very peripheral nodes of trees of the kind used here, were the rest of the system to remain the same.)

### 5.3.3 A Closer Look at Selected Trees

Three generalisation trees are now looked at in some detail to further illustrate the kinds of consideration that went into the design of such trees in general.

First I present the tree for generalising vocalic right contexts of fricatives, nasals and laterals (figure 5.3). The root-node or maximal level of generalisation is simply VOWEL, and the main split in the tree is between back and non-back vowels. Within non-back vowels, the primary split is between high front vowels and others. Further details may be gleaned from consideration of the tree. As examples of use, given sub-quorum quotas for [f_AXA] and [f_oeA] [4], data for the two classes (if any) would first be merged to form a class designated as [f_CV1A] [5]; if this class too proved non-viable, a further merge would take place with data for [f_laxA] (offset of [f] before the vowel of "bird") to form [f_CV2A] (offset of [f] before [ax], [oe] or [lax]), and so on. Note that [y] is treated as a vowel for the purposes of context-generalisation in this tree.

It may be noted that the tree used to govern generalisation of post-consonantal vowel-onsets — where generalisation is effected via the vowel itself rather than its context — is almost identical with that just shown. The presence of schwa

[4] [f_AXA] is the offset of an [f] before either a schwa or an [axD1] (first subphone of the diphthong in "go"), schwa and [axD1] being subject to immediate merge; [f_oe] is the offset of an [f] before an "open" [e], as might be found in a production of the word "fell" for example.

[5] "CV1" is meant to suggest something like "general central vowel (1)".

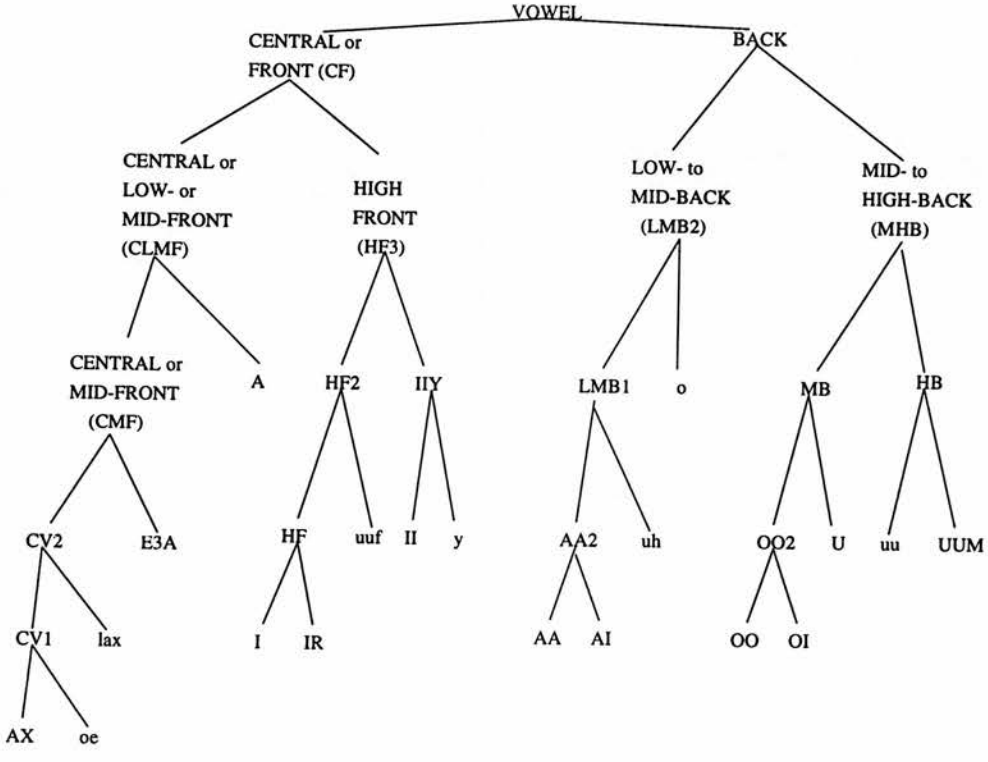



**Figure 5.3.** Generalisation Tree Schema for Vocalic Right Contexts of Offset-Subphones of Fricatives, Laterals and Nasals

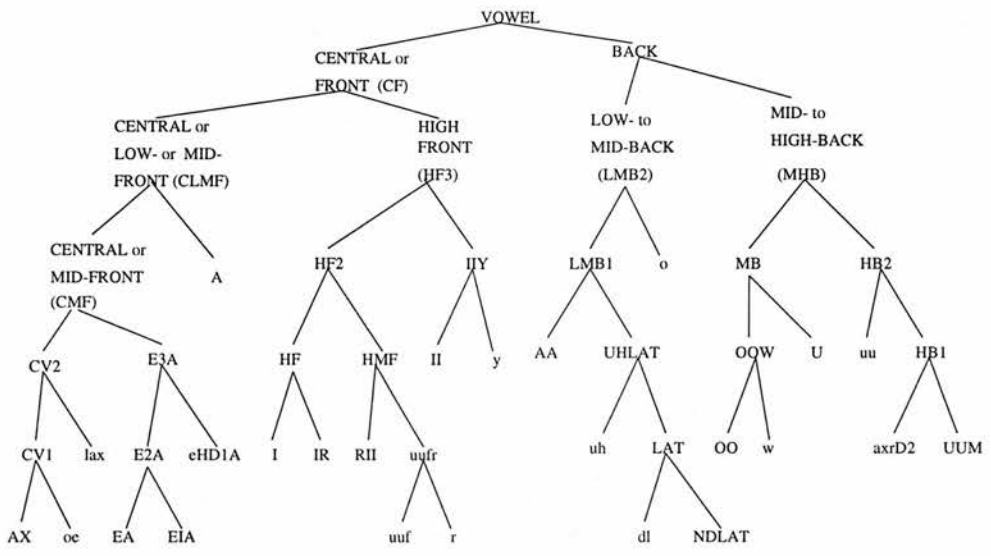



**Figure 5.4.** Generalisation Tree Schema for Vocoid Left Contexts of Onset-Subphones of Fricatives, Affricate-Releases, Laterals and /y/

([AX]) as a node in these trees is not ideal, because of the additional effect (here ignored) of the phone *following* schwa on its acoustic realisation (2.5.9), so that for example onsets of [lax] and [oe] with left consonantal context X get merged (where quora require it) with onsets of [AX] with left consonantal context X; this clearly fails to take account of the fact that the right context of schwas may cause their onset to show formant-movement in a particular direction from the outset, and it seems highly probable that recognition of the likes of [lax] and [oe] must suffer in consequence.

A second example is provided by the structure used to govern generalisation of left vocoid contexts of fricatives, affricate-releases, laterals and [y] (figure 5.4). This tree is similar to the previous one in many respects, but one of the differences arises from the desire to accommodate a few rare cases of consonants with [r], [w], or [y] as left context (in most forms of Southern British English this would always imply rapid speech with the disappearance of an intervening vowel). (Some of the node-labels do not reflect the inclusion of such cases in the tree, however, the node-names usually reflecting rather the dominant use of the tree, namely the generalisation of vocalic left contexts.)

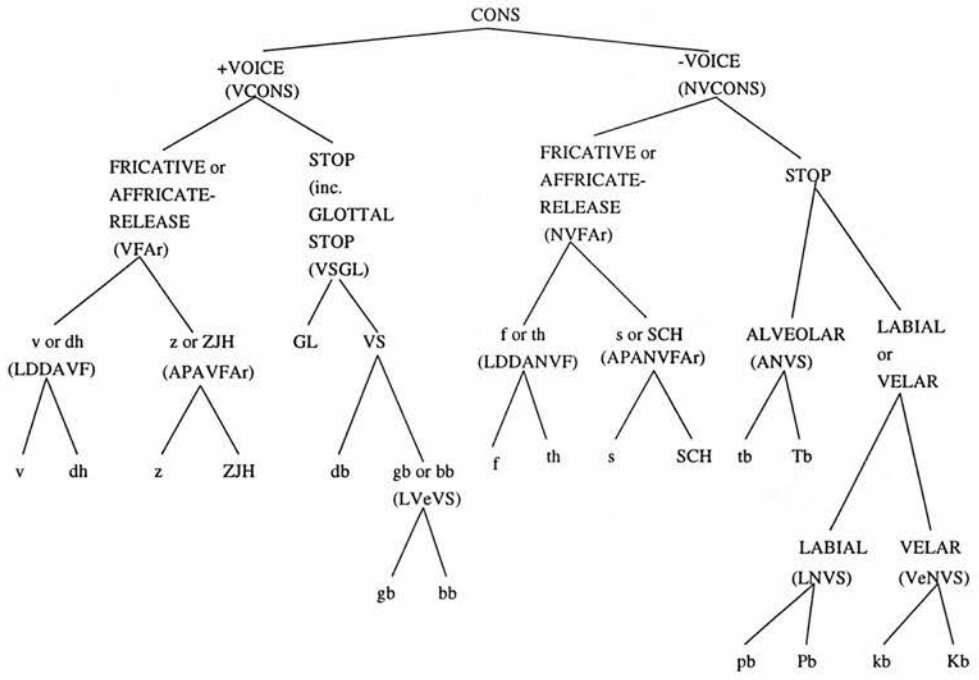



**Figure 5.5.** Generalisation Tree Schema for Consonantal Left Contexts of Onset-Subphones of Dark Laterals

Finding room for the consonants was not easy, apart from [y]. [r] is placed as sister to [uuf] in virtue of low F3 being a common feature, and laterals (almost always, one would think, dark laterals) next to [uh] in virtue of similar F1 and F2 (for dark laterals at least). [w] is impossible to place satisfactorily, and it is placed as sister to [OO] as the least undesirable option. The cases of these consonants immediately preceding consonants are in any case quite rare.

A final example is provided by the tree for generalising left consonantal contexts of dark laterals (figure 5.5).

It hardly needs stating that, beyond a certain point (which may come quite early in some trees), the less generalisation proceeds toward the root-level of a tree, the better. This is partly because it is not the case that we get a uniformly more diffuse distribution of vectors with each ascent to a higher level of generalisation — in some cases an ascent by one level may lead to a distribution only

slightly more diffuse, while in others (particularly at higher levels of generalisation) the increase may be very great (we could talk of natural, or more natural, clusters in the one case, and forced, or more forced, clusters in the second). In the generalisation of right subphones of pre-consonantal [+VOICE] fricatives, for example, it is certainly the case that the amount of information lost by merging [$\alpha$_mA] and [$\alpha$_bcA], or [$\alpha$_ngA] and [$\alpha$_gcA], is relatively small, and also that the amount of information lost by merging [$\alpha$_LVSNaA] ($\alpha$ with a labial [+VOICE] stop or labial nasal as right context) with [$\alpha$_VeVSA] ($\alpha$ with a velar +VOICE] stop as right context) is smaller than that lost by merging both of these with [$\alpha$_AVSNaA] ($\alpha$ with an alveolar [+VOICE] stop or nasal as right context). Once generalisation is forced to proceed beyond that point — merging, for example, all [+VOICE] obstruent and nasal right contexts, and even merging these with liquids and glides — distributions are then being formed from groups of data which can be quite dissimilar, so that modelling with a single gaussian in particular will result in poor representation. Clearly, in the extreme case we get back almost to context-independent subphones, undoing almost all of the work done to get discriminating models.

The generalisation-tree structures should not therefore be seen as making statements about the desirability of merges, or even simply about the relative undesirability of merges, but rather as reflecting also a desire not to lose necessary classes altogether. In constructing them, at the higher levels of generalisation the consoling thought was always, "Well, in a serious application of FURIDA, data would not be so short as to require such a merge, but in the meantime, the possibility needs to be accommodated".

### 5.3.4 Problems of Bias

Attention must be drawn to a particular weakness of the generalisation-procedure (or at least of the procedure when used in conjunction with modelling using single gaussians), which is that with unequal amounts of data for different specific classes, degrees of generalisation in the classes created vary considerably, resulting in some cases in a significant bias in favour of classes resulting from lesser degrees of generalisation.

As an example of the problem, consider the case of offsets of [z] before [pc], [kc], and [tc], where devoicing is almost guaranteed, so that acoustically there is often little or nothing to choose between these subphones and respective offsets of [s] before [pc], [kc] and [tc]. It happens that [s] is more common in the training-data than [z], so that there is every likelihood of being able to train specific models for [s_pcA], [s_kcA] and [s_tcA], while with [z] the more limited data may mean that [z_pcA] and [z_kcA] and even [z_tcA] may have to be merged together to get a quorum, forming a class [z_NVSA] (offset of [z] before a [-VOICE] stop) which will represent the three original classes rather poorly using a single gaussian. In recognising speech, given the acoustic similarity of real [s_tcA] and [z_tcA], and similarly for the other pairs, it is probably inevitable that a real [z_tcA] subphone will score more highly for the more accurate [s_tcA] *pdf* than for the [z_NVSA] *pdf* that represents it, and likewise again for the other pairs.

The best solution to problems of this kind would probably be a massive increase in the amount of training-data, to ensure that generalisation proceeded only a short way at most for any class, though some form of compensation for less well represented classes might be used in the absence of such an increase (McInnes *et al.* 1989a).

### 5.3.5 Defining Predecessor-Sets to Accommodate Generalised Classes

The existence of classes at different levels of generalisation, with their precise forms impossible to predict before training is complete, means that a flexible algorithm is required for defining class predecessor-sets, that is, for defining for each class that appears in the final inventory of classes the list of classes that may appear immediately preceding it in a phonetic transcription (4.4) (strictly speaking, of course, it is not classes that appear in sequence in phonetic transcriptions, but symbols representing classes, but I shall take the liberty of using the more manageable locution).

Predecessor-sets are defined initially for all specific classes (where 'specific'

includes classes like [PAL_EB] [6] resulting from immediate generalisations), prior to conditional generalisation, the predecessor-set for each class listing the specific (or immediately generalised) classes that can precede it. This series of basic listings is used to generate predecessor-sets expressed in terms of the classes (specific or generalised) that actually exist after training and clustering. To aid exposition, I shall refer to the basic listings as *specific term predecessor-sets* and to the predecessor-sets required for constraining the DP search (4.4) — all of the terms appearing in these must clearly refer to classes present in the class inventory — as *working predecessor-sets.*

For a given class whose working predecessor-set is to be defined (let us call the class the *target class* for ease of reference), the first step is to determine whether it is a specific class (where 'specific' is as before) or a generalised one (where 'generalised' is also as before). If the target class is specific, the list of specific classes in its specific term predecessor-set is worked through, and if any has representation in the class inventory, it is added to the working predecessor-set of the target class; for any that does not have such representation, the appropriate generalisation-tree is consulted (that in which the specific class figures at a leaf-node), and the nearest ancestor that does have representation in the class inventory is added to the working predecessor-set of the target class. If the target class is not specific, the appropriate generalisation-tree is consulted to find all the specific forms of the target-class (leaf-node descendants of the node at which the target-class figures), and each one of these is treated as a target-class in turn (that is, the procedures just described are followed for each one).

In the case of vowels, the procedure just described led to a number of complications as a result of the different directions in which generalisation takes place (on the vowel itself in some instances for vowels, on the context for consonants), given that there is no look-ahead beyond the immediate predecessor in the definition of predecessor-sets. As an example, the offset of an [r] preceding a vowel [aa], which is represented by the subphone [r_AAA], is a legal predecessor of the generalised subphone [R_LMB2B] (onset of a low mid-back vowel preced by an [r]); this generalised subphone in turn is included in the predecessor-set of [Co],

[6] onset of [e] or [eD1] following a palatal consonant

so that a subphonic transcription such as

[ r_AAA R_LMB2B Co ... ]

could be output, which clearly does not make good sense.[7] A surprisingly simple solution to this problem[8] is to introduce a clear separation between *classes* (as the objects for which sequence-constraints are defined, and from which paths are built in the transcription procedure) and *models* (the objects for which probability scores are calculated). The classes may remain everywhere specific, and generalisation can be restricted to models. When a model is generalised (e.g. we may estimate a single set of statistical parameters for a phone in three different contexts) several classes share it or are represented by it, but all sequence-constraints can be expressed in terms of specific classes. Unnecessary duplication of work in probability-scoring is prevented by calculating a probability-score just once for each model, and propagating scores to all its participant classes.

I introduced such a separation between classes and models in the case of vocalic subphones, in order to avoid the difficulties described above, and did likewise for TR classes and for stop-closures, but was prevented by shortage of time from introducing the separation across the board.

## 5.4 Standard Approaches to Clustering

The generalisation-procedure described above is a form of hierarchical clustering (Duda & Hart 1973), in that if two samples are grouped together at some level of generalisation, they remain so at all higher levels of generalisation. The procedure used here is, however, controlled by using manually constructed tree-structures, and it is common practice — and preferable, where it is possible — to employ automatic or nearly automatic procedures to determine which samples should be merged with which, and in what order merges should proceed (Lee *et al.* 1990b; Young & P.C.Woodland 1993).

[7] The subphonic transcription would be placing in sequence the offset of an [r] preceding a vowel [aa] as in "park" or a diphthong [ai] as in "aisle", the onset of an [aa] or [ai] or [uh] (as in "cup") or [o] (as in "not") following an [r], and a core [o] subphone.

[8] It is simple once it is pointed out, but is one which I failed to see for myself. I am grateful to Fergus McInnes for pointing it out to me.

There are two principal forms of hierarchical clustering, *agglomerative* or bottom-up procedures, and *divisive* or top-down or splitting procedures. In agglomerative clustering, one starts from specific classes and forms clusters as required, while in divisive clustering one starts with all the specific classes pooled together and works toward a partition into more specific classes. Forms of both have been used with success in ASR applications.

### 5.4.1 Agglomerative Clustering

An example of an algorithm for general agglomerative clustering is provided in (Young 1992). HTK clusters states of HMMs rather than entire HMMs, tying the state distribution parameters of clustered states so that the corresponding HMMs share a single set of parameters for the tied states (Woodland & S.J.Young 1993) (cf. a similar strategy used in CMU's Sphinx II recogniser (Hwang *et al.* 1993)). Thus the mental translation required for applying the clustering technique to the recognition units used in this work is very straightforward. The basic form of the algorithm is as follows:

Given that one wishes to generalise into a pre-determined number N of clusters,

1. Create a single cluster for each of the original class samples.
2. Set $n$ equal to the current number of clusters.
3. While $n$ is greater than the desired $N$, find the clusters $c_i$ and $c_j$ which are most similar, and merge them.

A variety of distance-measures can be used in determining the nearest pair of clusters. One measure used by the HTK researchers was based on the divergence between the two gaussians used to model the initial states or clusters, which for diagonal covariance matrices is given by

$$D(i,j) = \left[\frac{1}{d}\sum_{k=1}^{d}\frac{\sigma_{i,k}^2}{\sigma_{j,k}^2} + \frac{\sigma_{j,k}^2}{\sigma_{i,k}^2} - 2 + (\frac{1}{\sigma_{i,k}^2} + \frac{1}{\sigma_{j,k}^2})(\mu_{i,k} - \mu_{j,k})^2\right]^{\frac{1}{2}}, \tag{5.2}$$

where $\mu_{i,k}$ is the mean for the $k'th$ variable in state $i$, $\sigma^2_{i,k}$ is the variance for the $k'th$ variable in state $i$, and $d$ is the dimensionality of the representation vector (Young & P.C.Woodland 1993).

Agglomerative clustering should work best when the initial samples (prior to any clustering) are large enough to have a good chance of being representative of their classes. For, if a cluster ABC of samples A, B and C for three specific classes is to be a 'rational' cluster, and work well in making possible recognition of any a's, b's or c's as examples of the class represented by ABC, it is clear that the three sets of properties of the data (i.e. of samples A, B and C) on the basis of which the cluster ABC was formed should be genuine properties of the *populations* represented by A, B, and C. On the other hand, this becomes less important the less difference there is between the three populations: where the differences are insignificant, there is a better chance of modelling *each* of A, B, and C by means of ABC even if individually the class samples are not representative of their class, for the obvious reason that there is a fair chance that they will each misrepresent their class in different ways, and collectively do a fair job of representing each class reasonably well.

If this argument is sound, agglomerative clustering makes most sense in situations where comparatively large amounts of data are available, and one's motivation is not so much to find a way to overcome data-shortages as to find a principled technique for collapsing distinctions which are actually unnecessary for recognition (compare the first-line merges described for the likes of [$\alpha$_iiB] and [$\alpha$_iiD1B] above).

Once this has been said, it remains true that *human* judgement, too, when based on patchy data, may not be in a very good position to determine what is best merged with what, particularly when merges begin to be forced rather than natural. There may perhaps be room for debate about the possibility of extrapolation from the data immediately available using general phonetic knowledge, but it is probably wise to accept that since statistical speech recognition should only be attempted on the basis of adequate amounts of data, and since acoustic relationships between phonological categories can vary so much from speaker to speaker, automatic approaches to generalisation are preferable to manual ones in serious (practical) ASR applications (where data should be far more abundant

than in the present case). One question that then remains concerns the best form of automatic procedure to adopt. In the next section I look at another approach, based on automatic induction of decision-trees, that has tended to displace the use of agglomerative clustering techniques, partly because it provides for better handling of *unseen* classes — classes not seen in training but encountered in recognition-mode.

### 5.4.2 Clustering Using Automatically Induced Decision-Trees

A form of divisive hierarchical clustering based on automatic production of decision-trees has become popular in recent years (Lee *et al.* 1990b; Young *et al.* 1994). I shall describe the form the process takes in the article by Young, Odell and Woodland where it is again *states* of HMMs, rather than HMMs themselves, which are clustered, so that the potential for translation of the technique to a subphone-based approach can once again be easily seen (cf. again the clustering of states into *senones* in Sphinx II (Hwang *et al.* 1993)).

First, a list of categorical questions is drawn up by a phonetician or speech-expert of some kind, each question having the form "Is the left (or right) phone a member of the set X?", with X denoting a phonetic class which may be of any relevant level of generality (examples might be such classes as fricative, vowel, obstruent, [+VOICE], front, [ii], [s] and [t]). The drawing up of this list of questions is obviously not automatic, but clearly the questions are designed to cover all the distinctions which decades (centuries, perhaps) of work in Phonetics have shown to be likely to be of importance in the determination of particular contextual effects on a given phone. Once the pool of questions has been determined, an automatic process is used to determine which subset of the questions, and which sequence of the questions in that subset, is "best" for a particular set of candidate states (states that are candidates for merging), where a typical set of such states might be all final states of all triphones of the vowel [ii].

Let $S$ denote the set of all $n$ candidate states for some particular case of this kind. A tree is to be designed that will partition the states of $S$ into a smaller number of generalised states, so that states that are grouped together

into one of the generalised states may share a single set of spectral parameters. (HTK recognises speech using mixture gaussian *pdf's*, so there is an additional consideration in tree-induction: the data from all of the original states that get tied together at a terminal node must be sufficient to allow the estimation of the parameters of such a *pdf* (for some given number of mixture-components). But note that single gaussians are used for the initial candidates for clustering.)

The goal of the tree-induction procedure is to maximise the likelihood of the training-data given the tied states. Initially, *all* the candidate states for a given case are pooled at the root node (e.g. all final states of all triphone models for [ii]), and the log likelihood of the training-data is calculated. Each of the $N$ questions in the list of questions (or conceivably in a relevant subset from the list) is then tried out in turn to produce a series of $N$ different partitions of the pooled data into two, and the increase in log likelihood brought about by each such partition (over and above that for the pooled data) is then measured. The question which led to the partition yielding the greatest increase is taken as the best question and made the basis of the first split of the tree at root level. The process is then repeated for each of the two subsets of the originally pooled data, and iterated to grow further branches of the tree, until the increase in log likelihood arising from a further split falls below a certain threshold, or sample-sizes for states tied together at a terminal node fall below the minimum level required for the desired number of mixture-components. Finally, terminal nodes with different parents are tentatively merged, and the *decrease* in log likelihood this causes is measured; any pair of nodes for which this decrease is less than the threshold level used for stopping splitting are then merged. Young *et al.* report that this final move reduced the number of states by 10 to 20% without any adverse affect on performance (Young *et al.* 1994).

A number of points are worth noting about this procedure. Firstly, comments made above about agglomerative clustering apply again here: the algorithm can be expected to work best when the general level of data-sample sizes is high (this would remain true even if single gaussian models were all that were required); given that a threshold (quorum) is set, it follows that, for example, all initial states of [e] triphones with [s] as left context can only be clustered in a distinct cluster (cf [s_EB] in FURIDA) if there is sufficient data to allow this, so that if the

data is not sufficient some more generalised cluster (leaving all [s] and all [z] left contexts merged, say) will be the final cluster produced; the less data there is in general, the more generalised the clusters will tend to be, leading just as readily as the procedure described in 5.3 to unnatural clusters (splitting according to the most radical difference may still leave very different classes within each of the clusters). One could not necessarily argue in such circumstances that the clusters would be better because based on objective criteria, since with sparse data many training-samples may be unrepresentative anyway. Having said all this, with large amounts of training-data available this algorithm has much to recommend it. A second point worth making is that most of the work done by the tree-based procedure in HTK to tie together states of HMMs is already done as a matter of course in FURIDA, which models phones explicitly in terms of context-sensitive subphones — thus the result in HTK of inducing a decision-tree for final states of triphones of [ii] may well include a clustering of final states of (f)**ii**(dl), (s)**ii**(dl), (sh)**ii**(dl), ..., and the clustering thus produces in effect a class equivalent to the [ii_dlA] subphone. This, however, is not to deny that deciding which things should be identified (coalesced) on the basis of an objective criterion is preferable (subject to conditions about plentiful training-data rehearsed above) to deciding on the basis of what appears to make sense from mere visual inspection, and it is certainly the case that the use of 'piecewise' context-dependency in FURIDA was a decision forced by limited data; given the necessary data, it would no doubt be preferable to follow the algorithm described, or perhaps to start with a whole series of [ii_dlA] models, one for each left context of the [ii], and use a clustering algorithm such as this one to find the best compromise between "trainability and specificity" (Lee *et al.* 1980).

Before leaving the subject of automatic induction of decision-trees, it may be noted that (as with agglomerative clustering), a variety of criteria is available for determining the relative value of different questions at any node. In the paper by Hwang et al. cited above, for example, the best question is identified as the one which results in the greatest reduction in entropy. This is in essence another way of saying that the best question is the one which is maximally informative (Bahl *et al.* 1989) so that after the split we are taken as far forward as possible in the direction of the identities of specific classes previously merged into one;

thus, in a different setting, given a pooling of all objects of any kind whatever, a good first question might be "Is X animate?", or perhaps, "Is X the kind of thing that can be referred to with a count noun in English?". An obvious advantage of working with concepts such as entropy or divergence is that they are quantifiable, so that an objective measure can then be used to determine the precise value of any possible split.

## 5.5 Further Options for Dealing with Data-Shortages

So far in this chapter, the focus has been on an approach to overcoming data-shortages that involves collapsing distinctions between classes (accepting a reduction in the degree of context-specificity that can be afforded). A number of different strategies are also available, and are considered here briefly for the sake of completeness. Before reviewing these strategies, it will be convenient to consider further some basic statistical features of class-samples, and this occupies the next section. Returning to the kind of simple display of two-dimensional data presented in chapter 4, I first focus on refreshing or introducing intuitive notions of a number of key concepts which remain helpful when considering data in more than three dimensions, even though intuition in the sense of visualisation ceases then to be available. (This material is elementary, and may be skipped by any reader with even a modest knowledge of multivariate statistics).

### 5.5.1 Review of Elementary Statistical Properties of Multivariate Sample-Data

Visual cues for correlation are no doubt familiar, given a two-dimensional scatterplot of data-points. In normal data representative of its class, if there is zero correlation between the variables, the data-points fall in a single cluster of roughly circular or elliptical shape. Whether we get an elliptical or circular shape may depend on the scaling of the coordinate axes, but if the shape is elliptical and the variables are uncorrelated, the principal axes of the ellipse will be in alignment with the coordinate axes. (The first principal axis of an ellipse is the longest straight line passing through the centre of the ellipse and connecting two points

on its circumference, and the second principal axis is the shortest straight line that passes through the centre and connects two points on the circumference; the two axes are, of course, perpendicular to each other.) When two variables are correlated, however, the principal axes of the ellipse are out of alignment with the coordinate axes. A leftward-leaning orientation of the (first) principal axis indicates a negative, and a rightward-leaning orientation a positive correlation between the variables. The directions of the principal axes of the ellipse are determined by the eigenvectors of the covariance matrix, and their lengths by the corresponding eigenvalues. For present purposes it will suffice to understand the term eigenvector as equivalent to "direction along which there is greatest variability", and the term eigenvalue as the amount of variability along the corresponding dimension.

It should be noted that the discriminant function used in Chapter 4, derived from the *pdf* for the multivariate normal density, took account of whatever correlations there may have been between representation vector elements via the term $(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i)$, which gives the Mahalanobis or covariance-normalised distance from the vector $\mathbf{x}$ to the class mean vector $\mu_i$.

The significance of taking account of covariance may be highlighted by considering what effect it would have on the calculation of class-membership scores if we used Euclidean rather than Mahalanobis distance. In the two-dimensional case using Euclidean distance, all points having the same distance c from the mean vector satisfy the equation

$$(x_1-\mu)^2+(x_2-\mu)^2=c^2 \tag{5.3}$$

which is the equation for a circle with radius c about the mean vector. Given an appropriately scaled plot of two-dimensional data with equal variances and zero correlation, the Euclidean distance measure conforms perfectly with our intuition about relative distances of different data-points from the sample mean. Where there is zero correlation but unequal variances, Euclidean distance retains this property if the variables are first standardised by dividing the differences from the mean by the respective standard deviations. Where the data suggests a clear correlation, however, the inappropriateness of the Euclidean distance measure becomes clear. This may be illustrated by considering the direction of greatest

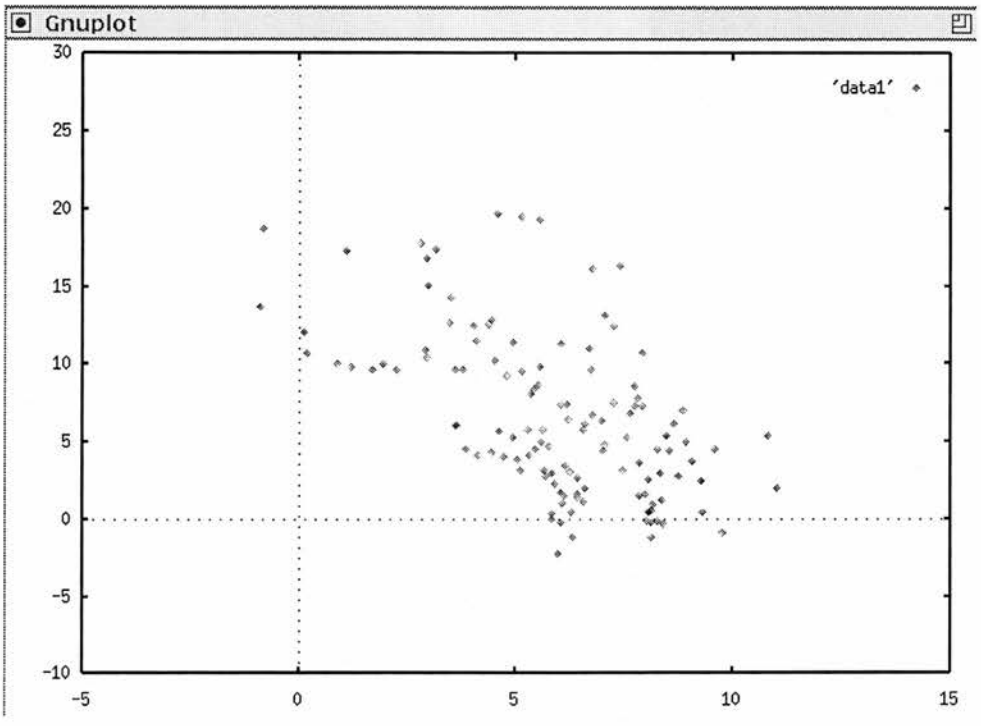



**Figure 5.6.** Plot of 1st static against 1st delta coefficient for PAL_MHBB

variability in the data plotted in figure 5.6, and imagining a data-point A at one extreme along that direction. In terms of Euclidean distance, a further data-point B positioned as far from the centre as A is, but at a point along the second principal axis, would be as close to the mean as A is. Yet our intuition — if we take the correlation as being germane to the class — would be that the data-point B was actually further from the class mean in some sense than was the point A, even though both lay on a circle centred on the mean vector. If we think rather of all points with constant Mahalanobis distance from the mean, we find that they satisfy the equation for an ellipse and if we were to inscribe ellipses of constant Mahalanobis distance from the mean over the data in figure 5.6, we would find that the data-point B fell outside the ellipse on which A was situated; we would have taken account of the correlation in calculating the relative distances. (I am indebted to (Flury & H.Riedwyl 1988) for this example).

In moving from two dimensions to three, an ellipse becomes an ellipsoid, and

in higher dimensions a hyperellipsoid; whatever the dimensionality, the principal axes are given by the respective eigenvectors, and the variances along the axes by the corresponding eigenvalues.

### 5.5.2 Statistical Profile of the Class Sample Data

In this section I present a statistical profile of the data, with the focus on two main features, namely (1) the degree of statistical dependence between individual representation-vector elements, and (2) the extent to which the variance- and covariance-structure varies from class to class. These two features are fastened on because they are the key elements of two assumptions that can be made in order to allow statistical modelling with reduced training-data – the assumption of complete statistical independence (zero correlation) between all vector elements, which allows use of diagonal covariance matrices, and the assumption of a shared covariance structure for all classes, which allows use of a single covariance matrix estimated from data pooled from all classes.

The inverse of a diagonal covariance matrix $\Sigma$ is formed simply by calculating the reciprocals of the diagonal terms of $\Sigma$, so that there is no longer an algebraic requirement for $d + 1$ vectors (where $d$ is again the dimensionality of the representation vectors); that said, however, estimates of variances that rely on $d$ or fewer vectors are not likely to be always reliable, particularly where $d$ is relatively small. The advantages of using diagonal matrices are rather to be found in the computational efficiencies this makes possible (cf. 4.6.1): the evaluation of

$$(\mathbf{x} - \mu_i)^t \Sigma^{-1} (\mathbf{x} - \mu_i)$$

reduces to $2N$ multiplications and $N - 1$ additions (giving, for the example case considered in 4.6.1, 24 million multiplications in place of 114 million using full covariance matrices).

With the assumption of a common covariance structure, each variable $i$ has a constant variance, and the same covariance with all other variables, irrespective of class, with only the mean vectors remaining class-specific; hence the precise location of each class in the multidimensional space remains unique, but the way the data spreads out from the location defined by the mean vector is the same for

every class. The absolute minimum requirement in terms of training-data is then only that the total number of vectors minus the total number of classes should be larger than the dimensionality of the data (Dalby *et al.* 1989).

One part of assessing the assumption of a shared covariance structure for all classes involves examining the variances. How does the variance for the i'th vector element vary from class to class? Table 5.4 illustrates the spread of the variances across the various classes for each element of the representation vector (given in the order: 10 cepstral coefficients, log power, dynamic features for the first four cepstral coefficients and for log power). The table makes clear the difference between maximum and minimum values to indicate the range of values found; given the possibility of extreme values being atypical, the standard deviation provides a firmer guide to the amount of dispersion, which in most variables is quite considerable (the standard deviation must, of course, be looked at in the light of the mean value). As far as variances are concerned, then, the clear suggestion of the data is that assumption of uniformity across all classes does considerable violence to the truth.

Turning now to consider the covariances, some light is thrown on both the issues mentioned above — statistical independence (zero correlation), and a common covariance structure for all classes — by means of the data presented in figure 5.7. Given a $d$-dimensional representation vector, there are $\frac{d(d-1)}{2-d}$ covariances to be considered (so 120 for a 16 element vector), and figure 5.7 illustrates the dispersion across all classes in the values of the correlations between the elements of each of the 120 pairings.

Once again the degree of dispersion in the values for any one pair across all classes is in all cases considerable. While it is probably true (assuming a distribution close to normal for each pair of variables) that some two-thirds of all classes have correlations within the range $-0.3$ to 0.3, this still leaves a third of all classes with measures of correlation outside this range, while the extremes illustrate that for every single pairing of variables there are at least two classes with a very marked degree of positive or negative correlation. Both assumptions — of statistical independence and of a common covariance structure for all classes — are thus shown to do considerable violence to the data.

| element | mean | Std Dev. | Max. class | Min. class |
|---|---|---|---|---|
| 1 | 5.984 | 3.542 | 24.340 TR22 | 0.206 w_oA |
| 2 | 2.522 | 0.903 | 12.535 MHB_GLNVCA | 0.248 aD1_axD2B |
| 3 | 2.212 | 1.167 | 7.728 V_hA | 0.073 R_RIIB |
| 4 | 1.511 | 0.723 | 4.445 OI_AXgl | 0.120 w_AA2A |
| 5 | 1.115 | 0.479 | 3.652 BACK_PVdhB | 0.113 h0SX |
| 6 | 1.105 | 0.509 | 5.503 MHB_GLNVCA | 0.137 r_EIA |
| 7 | 0.896 | 0.309 | 3.737 HB2_sB | 0.140 h0SX |
| 8 | 0.701 | 0.226 | 4.111 r_EIA | 0.082 w_AA2A |
| 9 | 0.590 | 0.203 | 1.910 r_E2A | 0.134 o_VCORA |
| 10 | 0.499 | 0.154 | 1.806 ng_silhA | 0.114 lax_VCORA |
| 11 | 0.473 | 0.263 | 1.810 NVSAc | 0.005 n_LVeVSA |
| 12 | 32.902 | 15.337 | 115.050 zh_VA | 1.598 w_oA |
| 13 | 16.032 | 6.671 | 44.506 TR8 | 1.855 w_oA |
| 14 | 12.599 | 4.821 | 30.611 z_LDDANVFA | 1.093 w_AA2A |
| 15 | 8.102 | 2.587 | 23.138 s_mA | 0.787 w_AA2A |
| 16 | 3.433 | 2.592 | 18.546 trc | 0.022 MHB_nB |

**Table 5.4.** Dispersion of Variances Across All Classes

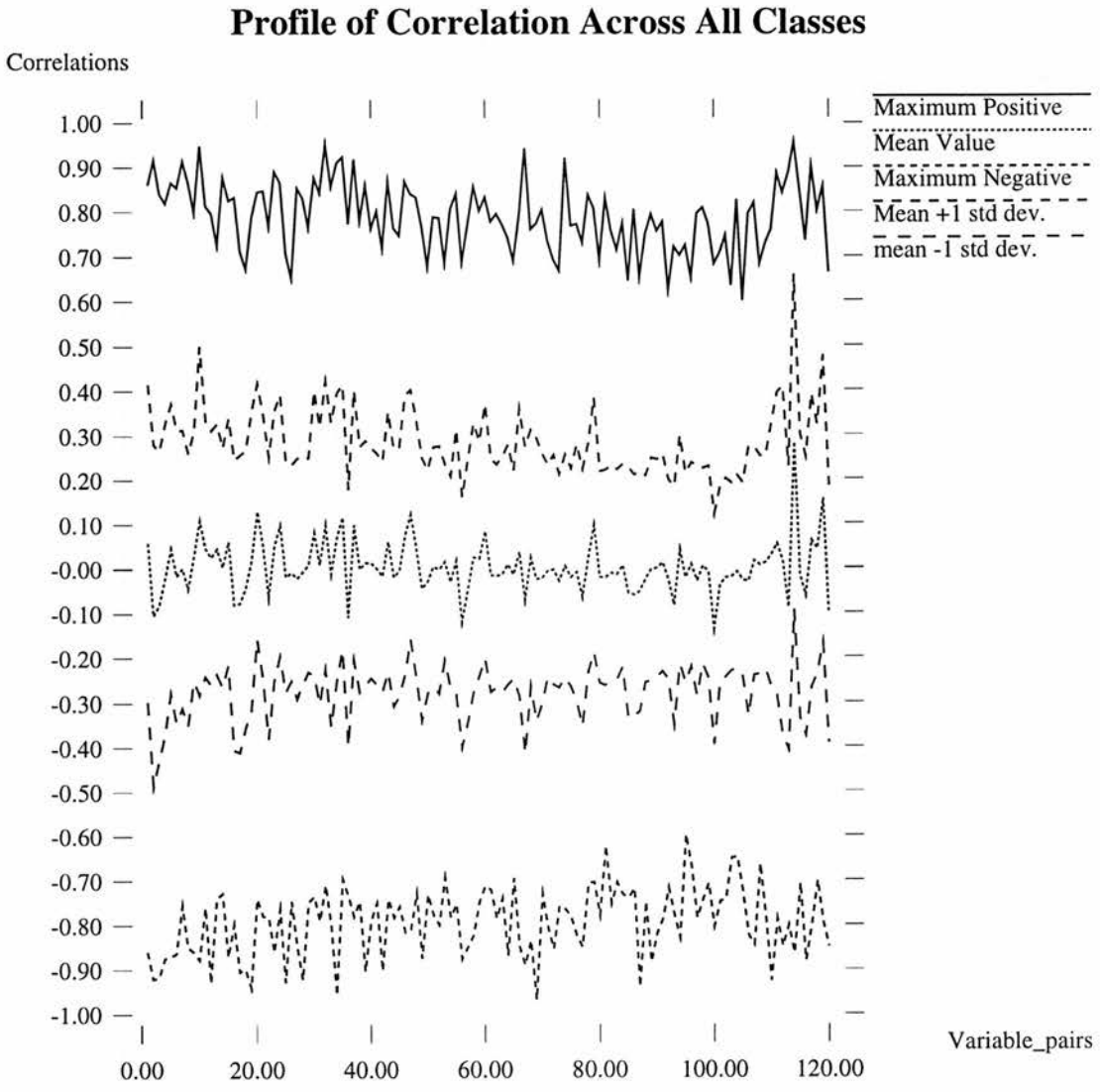



**Figure 5.7.** Dispersion of Correlations Across All Classes

### 5.5.3 Model-Fitting, Training Quotas, and Simplifying Assumptions

Modelling without making inaccurate simplifying assumptions about the data may involve estimating a large number of parameters [9], and large amounts of data may then be needed to make the estimates reliable. In the literature on statistical pattern recognition, it is frequently stated that making one or more inaccurate assumptions about the data to reduce the number of free parameters may sometimes lead to better results than can be obtained without such assumptions, if the training-data is insufficient for getting good estimates of the increased number of parameters required for more accurate models (Duda & Hart 1973; Bengio 1996; Bishop 1995). This, of course, is not to deny that the best results ought to be obtainable with accurate models trained on adequate amounts of data.

The case of parametric curve-fitting seems to be cited almost everywhere as providing a useful analogy with the case of statistical modelling. In Bishop's formulation of the curve-fitting example (Bishop 1995) synthetic data is generated by sampling a sinusoidal function $h(x)$ at equal intervals of $x$ and then adding to each sample-point a value taken randomly from gaussian noise with a standard deviation of 0.05. $h(x)$ thus represents the underlying pattern or systematic aspect of the data, which is obscured somewhat by the noise component overlaid upon it. The task is then to fit a polynomial $y(x)$ that will make good predictions for the values of new data generated by precisely the same mechanism. For this end to be achieved, a polynomial is required that that fits the underlying pattern represented by $h(x)$, rather than the precise points of the seen data. A linear polynomial (straight line) will fit rather poorly, while a high-order polynomial could be found which fits the seen data perfectly, and yet fails to predict unseen data well because it is fitted to the noise rather than to the underlying pattern. A lower order (cubic) polynomial can be obtained which provides an imperfect fit to the seen data but a fairly good fit to its underlying pattern, and so is able to serve rather well as a predictor of new data. Here then is a case where a model of intermediate complexity works best. The moral drawn is that the best generalisation to new data occurs when the number of free parameters is small

[9] Parameters that have to be estimated from data are known as *free parameters*.

relative to the size of the training set, and the moral is carried across unchanged to the case of statistical classification.

I do not wish to question the validity of the analogy, but there is perhaps a danger to be warned against. Given that one simplifying assumption is the assumption of diagonal covariance matrices, which reduces the number of free parameters from a hefty 136 to 16 for a 16 element vector (ignoring the mean vector for the present), one may be tempted to think of covariances as being somehow more 'noisy' or less germane to the data in some way, than are variances (not that Bishop is guilty of saying any such thing, of course). It is also true that the use of diagonal covariance matrices is widespread in ASR systems using continuous distributions for spectral modelling, which may also seem to point to covariances being of secondary importance (indeed it is often stated baldly that decorrelation is an automatic consequence of using a cepstral representation). In fact the common use of diagonal covariance matrices in ASR is probably prompted more by considerations of computational efficiency than anything else, since the use of mixture modelling — sometimes involving very considerable numbers of mixtures for each HMM state — makes the use of full covariance matrices expensive. The idea that covariances between cepstral coefficients are somehow more peripheral than variances, though, is a mistaken one, as the data presented earlier shows.

Where data is actually sparse, the danger is that it may seriously misrepresent its population, with serious consequences for classification-performance. *All* parameters are likely to be affected, the mean vector and variances just as much as covariances.

Given reasonable amounts of data, the assumption of statistical independence can be legitimated by the use of Principal Components Analysis (PCA), which can be employed to decorrelate data and may also make possible some reduction in dimensionality (Hadi 1996; Flury & H.Riedwyl 1988). PCA is not, however, a technique for coping with the effects of sparse data, and moreover involves making another assumption about the data — the sharing of a common covariance structure by all classes — which we have seen reason to regard as undesirable. Similar comments apply to another technique that can be used for obtaining a reduction in dimensionality, namely Linear Discriminant Analysis (LDA) (Flury & H.Riedwyl 1988; Duda & Hart 1973; Bishop 1995). Both techniques have,

however, been used to good effect in ASR, and while not suited to situations where data is patchy, can *reduce* the amount of training-data needed for a certain level of performance (Bocchieri & G.R.Doddington 1986; Haeb-Umbach & H.Ney 1992).

### 5.5.4 Pooling and Thresholding of Variance Estimates for Small Samples

Another strategy for coping with sparse data — it is little more than a crude fix — is to threshold the variances, or alternatively to pool variances, for small samples. One possible consequence of unrepresentative samples is that the variance-estimates are too small (they may also be too big, though underestimates are probably more damaging than overestimates), and setting a floor to variance-estimates for small samples, or pooling variances, is one way to cope with this problem.

When diagonal covariance matrices are used, the amended variances are simply substituted for the original values. Where full covariance matrices are used, merely substituting the ammended variances in the covariance matrix destroys the matrix's integrity, so that it is no longer guaranteed to be positive definite (and so no longer guaranteed to be invertible). However, it is at least not impossible that correlations estimated from small samples are likely to be more reliable than the variances, given that the ratio of vectors to subphones represented in the data for a particular class will be relatively low. Correlations between vector elements reflecting formant-movements within transitions may thus be less liable to misrepresentation than variances when samples are small. If this were to be so, adding further data to the data-sample might well lead to more drastic modification of the variance estimates and the means than of the covariances. On the assumption that this is so (and it is no more than an assumption) the substitution of ammended variance estimates while retaining the original eigenvectors may be a reasonable strategy (if crude fixes are to be bothered with at all). This may be effected by first deriving a correlation matrix from the original covariance matrix $\hat{\Sigma}$, and then recomputing $\hat{\Sigma}$ from the correlation-matrix and the ammended variances. The new covariance matrix is positive definite, and

preserves the correlations of the original ML estimate. The effects of using such a ploy have not been fully investigated here, however.

# Chapter 6

# Evaluation and Experimental Results

## 6.1 Introduction

It is becoming less and less common to find ASR systems attempting phonetic classification as an end in its own right (perhaps as a stage from which to progress to lexical access).[1] The majority of systems integrate phoneme recognition within the overall process of word- and utterance-recognition in a way that is probably closer to what humans do in making sense of what they hear, with lexical and usually also other higher-level constraints playing a major role in the recognition-process. In the system described in this work, the task is brought to an end with the production of phonetic transcriptions. This in part reflects the fact that a major goal of the thesis was to try to achieve automatic phonetic transcription as an end in its own right, but the decision to terminate the process with a *single* phonetic string reflects simply limitations imposed by time. If it is accepted that perfect phonetic transcription will remain an elusive goal, it follows that for viable

[1] A partial reversal of this trend may be in progress in one area, that of Information Retrieval from archives of speech material – it appears not impossible that an errorful phonetic transcription may provide a better basis for retrieval than an errorful textual transcription, since matching of query and "document" may take place via matches with parts of words (Ng & V.W.Zue 1997). It is also true that the scale of dictionaries that would be required for recognition of unrestricted material makes word-based retieval impractical in many instances, given current limitations on processor speed and on memory.

recognition from the basis described in this work, a lattice of scored phonetic hypotheses, rather than a single transcription-string, would almost certainly be required (McInnes *et al.* 1990). The only obvious possible alternatives to the use of such a lattice would be to use 'a priori' phonetic distances between classes, or, better, confusion-probabilities estimated from trials, to attempt lexical access from the single best transcription-string. (It seems likely, though, that concrete scored hypotheses from a lattice would provide a sounder basis from which to proceed.)

Given that termination occurs with the output of a single string, how is the performance obtained using the technique to be evaluated? And given that the system is complex (in the plain sense of having many parts and being the result of many design-decisions) how are its component parts to be evaluated (how far can we disentangle the effects due to different parts or features of the system?) in order to be able to identify the places where improvements should be made?

Evaluation of the system's performance in the first, general, sense is problematic in a number of ways. In so far as the concept of evaluation implies measurement against a standard, a basic problem is that it is not clear what standard to take. It is not obvious, for example, that human performance can be put forward as a basis for comparison. Even if it could, it is perhaps not clear what form of human activity should be tested to provide such a basis: recognition of excised "phoneme-sized" chunks of speech does not appear to be the right activity, because the system described here is at least given the benefit of having to find a transcription which fits over a whole utterance. (For what it is worth, Zue *et al.* report success-rates of only 60 to 70% for humans in the task of recognising "phonemes" excised from a database of several speakers, with minimal contextual information (Zue *et al.* 1989).) In any case, humans tend to abstract from "sub-phonemic" detail in the utterances they hear, "hearing" the message they think was intended rather than paying attention to the precise phonetic peculiarities of the utterance, so that it is difficult to get people to "hear" only the precise phonetic form of what is said, resisting an inevitable process of reconstruction into more useful linguistic currency. Moreover, in so far as humans do perform phonetic recognition, they probably employ all the well-known higher constraints that are absent from the present system. So the system's performance does not

seem to be comparable with human performance (in any sense of the word "comparable"!). As far as comparison with other automatic systems is concerned, most existing systems are judged on the basis of words recognised, rather than of phones recognised, and though there are published figures for the performance of automatic phonetic classification-systems (the last work referred to provides one example), evaluation of the current system by comparison with these would have required that things be rigorously set up to ensure a fair comparison, which in turn would have had an undesirably powerful effect on the evolution of the system, had it been made into an overriding priority. It was deemed more sensible to develop the technique in the ways that seemed most logical at each step along the way, and to give less priority to issues of comparative evaluation; the latter might never have become a live issue in any case, had the technique proved to be hopelessly ineffective. Consequently, evaluation in the sense of ranking alongside alternative techniques or systems will not be attempted, though figures will be given for performance which readers familiar with performance-measures for other systems will probably find it hard not to use as the basis of a "verdict" on the relative merit of the technique.

Termination with a phonetic (as opposed to full "orthographic") transcription, and with a *single* phonetic transcription-string, makes the measurement of performance difficult in other ways. Theoretically sophisticated evaluation of phonetic classification systems is possible, based upon entropy-measures, given a lattice and a language-model (McInnes *et al.* 1989b; McInnes 1991), but given the restriction to a single string and the desire to measure the *phonetic* accuracy of the transcription, what was rather required here was a scoring algorithm which compared manual and automatic transcriptions in a phonetically intelligent way, that is (in simple terms) in a way which took account of phonetic similarity between classes. Lack of time prevented the development of such an algorithm, however, and I was obliged to fall back upon available scoring software that is actually more ideally suited to word-recognition systems, using per cent correct and accuracy figures to assess the quality of phonetic strings output by the system.

The percent correct figure is defined as

$$\%correct = H/N \times 100\%,$$

where $H$ is the number of correct labels in the automatic transcription, and $N$ is the total number of labels in the manual transcription, and accuracy is defined as

$$accuracy = (H - I)/N \times 100\%,$$

where $I$ is the number of insertions. (An equivalent formula for the percent correct figure is $(N - D - S)/N \times 100\%$, where $D$ is the number of deletions and $S$ is the number of substitutions.) I used the HTK utility HResults to compute these two measures. HResults takes automatic and hand transcriptions for a given utterance and finds the optimal alignment between them using dynamic programming.

Some quite serious reservations need to be expressed about the use of HResults for this purpose. Firstly, while in the case of word-recognition (at least in typical speech-recognition tasks) there is very limited room for disagreement about what the correct word-sequence is, with phonetic transcription the room for disagreement about the correct or best phonetic transcription is greater. HResults is blind to "degrees of incorrectness", so that, for example, confusions between [ir] and [ax] or between [i] and [ir] are — to it — errors of precisely the same magnitude as confusions between, say, [oo] and [h] or [ai] and [s]. This blindness can sometimes cause HResults' "optimal alignment" to be far from optimal in reality, with potentially significant consequences for scoring. The following example of HResults output illustrates these points; first is shown the correct manual transcription of the sentence "The price range is smaller than any of us expected", and then there follows FURIDA's attempt at transcribing the utterance, aligned by HTK – the reader is asked to compare the first line of the manual transcription with the first line of the automatic transcription, and similarly for the second and third lines:

```
Aligned transcription: sc001.hand vs sc001.auto

     dh  ax    pc prb r      AIgl D3 s r     EIgl D3 n   zh  i z
s m      oo   l ax  dh  ax n e n ii v ax s    ax kc  kb s upc Pb
e  GLsc            tb ir dc db

TDHS dhR ax bc pc prb r aaD1 AIgl D3 s r eD1 EIgl ng jhc jhb ir z
```

```
s m ooD1 OIgl l lax dnc ax n a n i  w ax s tb y GLsc kb s
ax GLsc bb a kc tc tb ir dc db
```

Two of the diphthongs here — the [ai] in 'price' and the [ei] in 'range' — illustrate the point about discrepancies between transcriptions not necessarily being very significant (being differences that might well occur between two human transcribers of the same spectrographic data). The manual transcription for 'price' has only [AIgl D3] where the automatic transcription has [aaD1 AIgl D3], and so the latter is technically guilty of an insertion. In the case of 'range', the manual transcription has [EIgl D3] while the automatic transcription has [eD1 EIgl] and so is guilty both of an insertion and a deletion. The criteria for labelling of diphthongs stated in Chapter 2 (2.5.10) involve consideration of whether there appears to be a steady state before the diphthongal glide, and whether any final state is sustained, however briefly, at the end of the glide, and 'D1' and 'D3' elements are transcribed or not accordingly, but there are cases where the decisions are borderline. This is not to say that the distinctions or the criteria are meaningless or impossible to apply, but only that disagreements about particular cases are to be expected, so that where the disagreements are between manual and automatic transcriptions they may not be as significant as some other errors.

Again, errors such as the misclassification of [zh] as [jhc jhb] or of [dh] as [TDHS dhR] are of a different degree of seriousness from errors like the misclassification of [v] as [w] or the insertion of a [tb] in [s]. In the first group of errors, the differences are of a kind that would probably be insignificant in a word- or utterance-recognition system, where both forms might well be treated as acceptable variant pronunciations of a word or part of a word. HResults treats all substitution errors in the same way, however.

This same piece of HResults output illustrates the room for odd alignments when proceeding without phonetic eyes (so to speak). The automatic transcription of 'expected' is as [y GLsc k s ax b a k t ir d], with the unaspirated [P] being misrecognised as a [b], and a schwa inserted between the [s] and the closure for the [b]/[P], yet we get the following alignment from HResults:

```
HAND:  ax  kc  kb  s  upc  Pb  e  GLsc            tb  ir  dc  db
AUTO:  y  GLsc kb  s           ax GLsc bb a kc tc tb  ir  dc  db
```

If the alignment were rather

```
HAND: ax  kc  kb  s    upc    Pb  e  GLsc   tb  ir  dc  db
AUTO: y GLsc  kb  s ax GLsc   bb  a  kc tc  tb  ir  dc  db
```

it would provide a better reflection of the reality. Part of the problem here is that HResults, doing its best to find what it thinks is the "best" alignment, matches up the (final) [GLsc] in the automatic transcription with the [GLsc] in the manual one, failing to take any account at all of the phonetic distances between any pair of non-identical phones.

Unfortunately, an awful lot of weight is placed upon 'headline figures' (per cent correct and accuracy), and when systems are seen to be judged (and dismissed) *solely* on the basis of these figures, it is difficult for researchers not to succumb to pressure to concentrate on getting their headline figures higher at all costs, and even to cease worrying about whether these figures are actually the best indicators of underlying realities.

I succumbed to such pressure in taking some steps to reduce the extent to which differences deemed to be minor lowered the performance-scores. (It may be recalled that a number of distinctions were introduced in the first place to try to make the best of the Normality assumption.) The following are all the steps of this kind that I took. Firstly, the four differentiated 'D3' targets, [CiiD3], [CiD3], [CeD3] and [CaxD3] were collapsed to a single [D3] category in both the manual and automatic transcriptions at the time of comparison. Secondly, because of the absence of real spectral differences between the [aD1] element of [au] on the one hand, and [a] on the other, and similarly between the [axD1] element of [ou] and "canonical" schwa on the other, when either [au] or [ou] were realised pre-consonantally in a curtailed form as [aD1] *only* or as [axD1] *only*, they were treated as equivalent to [a] and [ax] respectively (cf a similar policy adopted at

the labelling stage for curtailed realisations of [ai] as [aaD1] and [oi] as [ooD1], as described in 2.5.10).

This chapter focuses on the evaluation of components of the system, rather than on comparative evaluation with respect to other systems. It needs to be acknowledged at the outset, though, that the degree to which components can be assessed "in isolation" varies from case to case. The features of the technique that are given particular examination are as follows:

- the handling of border-straddling frames via "TR" classes
- the completeness and effectiveness of the phonetic sequencing constraints
- the cost to the system of the lack of any explicit duration-modelling
- the extent to which the assumption of Normal distributions leads to loss of information
- the issue of representativeness of training-data
- the proportion of error that may be attributed to inadequate handling of pre-pausal effects
- the effect on recognition of variation in the frequency-warping and clustering adopted for derivation of the cepstral representation.

In addition, a number of experiments will be described that were concerned with attempts to address one of a number of pervasive problems that appeared in the output, namely the misclassification of certain monophthongal vowels like [aa] and [oo] as rising diphthongs ([ai] and [oi]).

It should be assumed throughout that except where it is explicitly stated to be otherwise, the framework for all experiments was as follows: 252 training-utterances were used (190 ATR sentences and 62 TIMIT sentences) and the performance figures given are normally for 200 ATR sentences classified in open test in twenty batches of 10, the open test condition being met by leaving out

from training the 10 ATR sentences to be classified in the given batch.[2] A particular cepstral representation was initially chosen arbitrarily from the set available, and only when most experiments had been completed was the system that had been found to be optimal (of those tested) then rerun with different cepstral representations to find which representation performed best. The initial representation was similar to that detailed in 3.4.3 as "Scheme C", involving less steep ascent to wider bands than in a pure mel frequency transformation, with bands non overlapping and with no tapering by means of triangular or similar functions. The representation comprised the first 10 cepstral coefficients, log power, and dynamic coefficients for the first four cepstral coefficients and for log power (as described in chapter 3). All classification tests used full covariance matrices.

In general, the better of any pair of configurations at stage $n$ in the experimental sequence was carried forward as a "control" or baseline for a new experiment at stage $n + 1$, only one variable or parameter-setting being changed each time. (In reporting below the experimental results at each stage, I generally repeat the figures for the control immediately before those for the new configuration, for easy comparison by the reader.)

For most of the experiments described, figures are given for confidence (c-value) and significance (P-value)[3] with respect to three diagnostic measures, namely the number of deletions, the number of substitutions, and the number of insertions[4]. In each experiment, interest focuses on the degree of any improvement in going from a control condition to a new or test condition, and hence on decreases in any of the three diagnostic measures being used. The absolute figures point to any such improvement in their own right, but there may be a

[2]This account glosses over, for the sake of simplicity, the more exact truth that one of the twenty batches comprised only 7 test sentences rather than 10, most of the data for a particular phonetic class being found within the three excluded sentences; it was thought better to make these three sentences available as training data, and preserve the open-test condition by excluding them from test-data in their batch.

[3]In most cases where such figures are not given, it is simply because the data needed for their calculation had been erased either through oversight or because of pressures on storage-space.

[4]I am indebted to Fergus McInnes for his help in connection with the subject-matter of this and the next two paragraphs.

question regarding the extent to which we can have confidence that a similar result would be obtained with different or with additional training and/or testing data; any difference between the results for the two conditions might, after all, reflect factors which are not relevant to the main issue (the result of "random variation" or of peculiarities in the sample (training-data or test-data)). Intuitively, we may be inclined to say that the larger any difference is (relative to the number of cases examined – here, the number of phones classified), and the larger the number of cases examined, the less likely this will be, but it may be felt that something more precisely quantified than this is desirable. Confidence- and significance-measures provide this, but it should be noted that their calculation can involve the making of assumptions that are dubious.

The c-value represents the posterior probability that the mean difference for a given feature is negative between the control and the test condition, and so points to an improvement over the control condition. It is calculated on the assumption of a prior uniform distribution for this difference (which is to say that all values for the mean difference between some limits are assumed to be equally probable, which in turn implies a probability of 0 for the mean difference being precisely 0). For a given diagnostic feature such as deletions, the c-value is calculated from the t-statistic for the matched sets of numbers of deletions, which in turn is calculated on the assumption that the number of deletions for any one sentence under a given condition is independent of the number of deletions in any other sentence under that condition, an assumption which may inspire some misgivings. In calculating the c-value, the degrees of freedom equal the number of paired sentences minus 1.

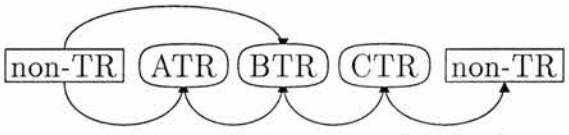
The more frequently used P-value represents the probability of the given difference occurring by chance were the null hypothesis of equivalence between the two conditions true. A significance level $P_0$ is usually taken as a threshold for rejecting the null hypothesis: if $P_0 = 0.05$ is chosen for example, we would reject the null hypothesis if the P-value fell below 0.05, the result being taken to be significant at the 5% level. The threshold adopted here is 0.017, arrived at by dividing 0.05 by the number of tests performed (one for each of deletions, substitutions and insertions); this adoption of a more demanding significance-level when using multiple tests is known as a 'Bonferroni manoeuvre'. The P-values given below

are derived directly from the c-values ($P = 2 \times c$ if $c \leq 0.5$ and $P = 2 \times (1 - c)$ if $c > 0.5$). The conversion-formulae reflect the fact that while the c-value is concerned with the probability of a particular direction of difference, the P-value is concerned with the probability of deviations in either direction from the null hypothesis of equivalence between the two conditions.

## 6.2 The handling of "TR" classes, Part I

In 4.6.2 I described how in the baseline system the TR classes built from border-straddling frames were handled in the transcription algorithm, and reported on a criticism and a suggestion for improvement made by F. McInnes. In this section I describe the results of a partial implementation of that suggestion (in 6.7 I return to the theme and report results for a full implementation).

In this first experiment the individual elements of TR "segments" are represented by distinct classes, so that the problems described in 4.6.2 may be avoided, but the resolution into distinct classes is merely nominal, all three elements of a TR class sharing a single model (a single set of cepstral parameters). The sequence-rules used may be gleaned from the following diagram:




This enforces the constraint that TR "segments" be at least two and at most three frames long, while the resolution into separate elements allows for an optimal location of the TR "segment" as a whole (4.6.2).

The results for two trials, one without the resolution into discrete elements and one with (all other things being held constant) are given in table 6.1 and table 6.2, with confidence and significance figures in table 6.3.

It is evident from the results that the resolution into discrete elements made a significant difference to the system's performance, with approximately a 2.5 point improvement for the percentage correct figure (about an additional 250 phones correctly recognised, some 200 fewer phones being deleted), though it is also evident from the figures for insertions (about 300 more with resolution of TR

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 66.96 | 50.98 | 6718 | 582 | 2733 | 1603 | 10033 |

**Table 6.1.** Monolithic TR Segments

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 69.43 | 50.5 | 6966 | 385 | 2682 | 1899 | 10033 |

**Table 6.2.** Nominal Resolution of TR Segments

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 1.0 | 0.0 |
| Substitutions | 0.94626 | 0.107148 |
| Insertions | 0.0 | 0.0 |

**Table 6.3.** Confidence and Significance Measures

class elements) that having separate TR elements made it easier for the system to posit changes from one phone to another and so to "oversegment". I suppose this is a consequence of the fact that with resolution of TR class elements the transcription-algorithm had more opportunities to "see" TR segments, and more opportunities to find pairs of phones that "fitted" to either side of them. In section 6.7 I return to this area and to an attempt to redress this problem.

## 6.3 Completeness and Effectiveness of Sequence Constraints

In 5.3.5 I referred to a proposal by F.McInnes that a clear separation be made throughout between classes and models, with models rather than classes being made the subject of generalisation, and specific classes thus sharing a single generalised model where necessary. This arrangement, it was pointed out, allowed sequence-constraints to be stated in wholly specific terms and thus opened up a way to overcome a particular problem I had faced as a result of the unusual way I had organised generalisation involving vocalic subphones. Unfortunately time precluded my implementing this arrangement for all classes across the board, and as a result the amount of constraint operating upon the sequencing of phones in the transcription algorithm is significantly less than it could be. In this section I wish first of all to demonstrate that this is so, and then to consider how significant an impact there might be on recognition-performance were the separation of classes and models to be fully implemented.

In the system described in this work, sequence-constraints are defined only across the narrowest possible 'window', that of pairs of subphones whose possibility of co-occurrence in sequence is at issue. The higher-order sequence-constraints that are the subject-matter of *phonotactics* (if we construe the term widely to include not only within-word and within-syllable restrictions but also restrictions that apply across word-boundaries) require a wider context to be taken into account. For example, while the pairs [th f] and [f dh] are permissible across a word or syllable-boundary, the triple [th f dh] is not. Now most illegal triples of consonants could in principle be blocked even within the limitations of the current

system and even without comprehensive separation between classes and models ("CSBCM" henceforth for short). The qualification "in principle" is necessary because of the fact that the sequence-rules that are available to block them are dependent for their effectiveness upon data-shortages not forcing generalisation up to levels where relevant distinctions are collapsed. (That the qualification needs to be made is in itself an acknowledgement that CSBCM would be preferable). An example will make this clearer.

Returning to the case of [th f dh], let us focus on the subphones of the middle consonant, which with copious data might be wholly specific even without CSBCM, i.e. [th_fB] and [f_dhA]. Given such subphones in the class inventory, the illegal triple is blocked by the non-appearance in the sequence-rules of

$$* \text{ [th\_fB]} \longrightarrow \text{[f\_dhA]}$$

(I use a leading asterisk to indicate that the sequence is illegal, but shall continue to use asterisks within phonetic labels as 'wildcard' symbols). The consonant [f], like every other English consonant, is able to appear as the middle consonant of only a restricted set of triples of consonants. [f] may follow only [m] ("bumph") or [dl] ("elf") at the end of a word or syllable, and when it does so, clearly any consonant (excepting [ng]) may follow it. After any consonant other than [m] or [dl], the [f] would have to be syllable-initial, and in that case the only consonants that may follow it are non-dark or syllabic laterals ("flea", "fallacious") or [y] or [r] ("few", "freckles"), or possibly syllabic nasals ("familiar", "funicular"). Each consonant of English similarly has a number of restrictions in respect of three-consonant strings it may appear in the middle of, and it is possible to integrate these restrictions even in a system like the present one which never looks beyond the immediately next or last subphone, by regulating the within-consonant subphonic sequences.

The only reservation that needs to be expressed about the relevance of these sorts of facts is that they are primarily phonological, and concerned rather with the abstract internal representations we have of our language than with what may actually occur at the level of physical utterance. The problem is just that it is possible in practice, for example, for an utterance such as "For the love of Heaven!" to be produced without the speaker successfully achieving phonation

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 69.43 | 50.5 | 6966 | 385 | 2682 | 1899 | 10033 |

**Table 6.4.** Nominal Resolution of TR Segments

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 69.33 | 50.71 | 6956 | 388 | 2689 | 1868 | 10033 |

**Table 6.5.** With Additional Sequence-Constraints

between the initial [f] of "for" and the [dh] of "the". In a majority of cases even short schwas do manage at least a pitch-pulse or two, but cases of devoiced or "deleted" schwa are by no means rare. However, when all this has been said, given the relative infrequency of "deletions" of this sort, more is probably to be gained by the inclusion of sequencing-constraints of the kind under discussion than is to be lost.

As an experiment designed to see how valuable these consonant-internal sequence-constraints might prove, I added a complete set of such restrictions to the sequence-constraints used in the previous configuration, and ran the system again, all other things being held constant. The results are shown in tables 6.4, 6.5 and 6.6.
The gain made from adding the further constraints is obviously very meagre, if gain it is (ten fewer phones correctly recognised, but with 31 fewer insertions,

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.128939 | 0.257878 |
| Substitutions | 0.196908 | 0.393816 |
| Insertions | 0.998738 | 0.002524 |

**Table 6.6.** Confidence and Significance Measures

only the insertions result being significant). It is instructive to examine the likely reasons for the meagre impact of the additional constraints, and in particular to examine how far the failure to implement CSBCM prevents the sequence-constraints from doing all the work they are intended to do.

As an example, consider a fairly extreme situation where data is so short that all offsets of pre-consonantal [f]'s are generalised to [f_CONSA]. In such a case, even though constraints of the form considered above may be present in the system, they cannot have any effect. We cannot, for example, impose the legitimate constraint that after [s_fB] only [f_yA], [f_rA], [f_NDLATA], [f_dlA], [f_nsylA], or [f_msylA] are legal (ignoring vocalic successors), since none of these exist in specific form. In the present system (without CSBCM)[5], the sequel comprising legal successors of [s_fB] is built up by taking each of [f_yA], [f_rA], etc. in turn and finding their nearest represented ancestor in the relevant generalisation-tree, which for all of them in this case is, we are supposing, [f_CONSA]. Given that we must allow

$$[\text{s\_fB}] \longrightarrow [\text{f\_CONSA}],$$

it is plain to see how we end up also allowing

$$[\text{s\_fB}] \longrightarrow [\text{f\_CONSA}] \longrightarrow [\text{NVCONS\_thB}]$$

(for example), which allows the phonotactically illegal string [s f th] to appear in output transcriptions.

Were CSBCM to be implemented, phonotactically illegal strings of this particular kind could be prevented from appearing in transcriptions. Without CSBCM, such strings could still be prevented given copious amounts of training-data (minimal generalisation being required). As an alternative to CSBCM, one could attempt to redesign generalisation-trees in a way that took more account of phonotactically sensitive areas, as far as possible tailoring generalisation-trees for each consonant to avoid as far as possible (or postpone for as long as possible) generalisations which collapse distinctions that are important for phonotactic constraints to be able to bite. An example of a particularly unwise design-decision

[5] comprehensive separation between classes and models

is provided by the way I organised the generalisation of left contexts of [n] and [m]. I grouped nasal left contexts of these phones together with vocalic ones, failing to anticipate the fact that while

$$[\text{VOWEL\_nB}] \longrightarrow [\text{n\_*A}]$$

and

$$[\text{VOWEL\_mB}] \longrightarrow [\text{m\_*A}]$$

hold,

$$*\ [\text{m\_nB}] \longrightarrow [\text{n\_*A}],$$

and

$$*\ [\text{ng\_nB}] \longrightarrow [\text{n\_*A}],$$

and

$$*\ [\text{n\_mB}] \longrightarrow [\text{m\_*A}],$$

and

$$*\ [\text{ng\_mB}] \longrightarrow [\text{m\_*A}],$$

certainly do not. Given the lack of data for nasals with nasal left contexts, generalisation of left contexts proceeds in all cases to a level where the distinction between nasal and vocalic left context is collapsed into a common identity, with the result that strings of the form [m n f] are able to appear in transcriptions. This could be prevented if suitable alternatives to vowels could be found to partner nasals as left contexts of [m] and [n] (not necessarily an easy order to fulfil), or of course if data was so plentiful that no generalisation was needed at all, or of course via implementation of CSBCM.

It is worth mentioning also the particular weaknesses — as far as the potential power of sequence-constraints is concerned — of context-independent classes, and in particular of the stop-closures ('in particular' in view of their frequency of occurrence). I introduced SBCM for stop-closures, having decided to pool data

for specific stop-closures in the ways described earlier (2.5.2), but even after the implementation of separation between classes and models, the stop-closure classes are still such entities as [tc], [bcv], [tgc], and provide no cue to the identity of their left context. As with the consonant-internal constraints discussed above, there is in fact considerable scope for enforcing additional legitimate constraints in this area. Consider the closure for [p], for example. *Whatever* precedes a [p]-closure, the [p] may be released into a vowel or [y], [r], [l], or syllabic [dl] or syllabic nasal. After [s], [m] or [dl] a [p] may be syllable-final, so that the [p]-closure may be followed by a release into any phone other than [ng], (with the likelihood of non-release being greater for particular consonants). After any other consonant, however, a [p] has to be syllable-initial, and since in that case the only consonants that may follow it are [r], [l], [cl] and syllabic [dl] or syllabic nasals, we would again have the potential for useful constraints on the transcription process if we incorporated a contextual annotation into the closure-label to indicate its preceding context.

Whilst the theme of separation between classes and models is in the air, it may be useful to point out that there is not in principle any reason why the process of model-sharing by specific classes should be restricted to bare subphonic or phonetic categories. We could make further nominal distinctions between subphonic classes according to other features such as syllable-position, distinguishing for example between phones in each of their possible structural positions within syllables. This would allow us, moreover, to use a more sophisticated form of duration-modelling when tracing the most probable transcription using dynamic programming, We could, for example, take account of the relative durations of a hypothesised [ii] and a hypothesised [z] in a hypothesised rhyme, and let this hypothesis compete with another in which the [ii] was syllable-final and the [z] syllable-initial (cf 4.8). This would also make possible the integration of a significant amount of additional phonotactic constraint.

In conclusion, it does seem reasonable to claim that significant improvement in performance would result from the introduction of CSBCM (even if measures would be required for economising on search during the transcription process as a consequence of greatly increased numbers of classes — even with model-sharing and the propagation of scores to all sharing classes from a single score-calculation,

the transcription would become laborious without such measures).

## 6.4 Cost to the system of the lack of any explicit duration modelling

I argued in chapter 4 that naive duration modelling is unlikely to be particularly helpful because of the powerful influences on phone-durations of factors other than plain phonetic identity. In this section I describe an attempt to put this claim to the test by incorporating naive duration modelling in the transcription algorithm, and comparing the recognition-performance with that of a system-configuration that was identical except for the absence of duration-modelling.

The procedure used for implementing the test was in nearly all respects similar to that described in 4.6.3 for the particular case of stop-closure duration scoring. Durations for each category of phone — largely ignoring context — were collected from the training-data, and histograms built from them. Each bin represented a single number of frames, ranging from 1 to 35. The absolute numbers in each bin were converted to relative frequencies (the percentage of the total number of tokens that fell into each particular bin), and after some smoothing (using a simple 5-point moving average) the negative logs of the converted values were taken as appropriate penalties to apply to phone-hypotheses of the given length in paths traced in the transcription program. (In smoothing, any leading or trailing sequences of zeros were "protected", in the sense that they were not averaged with their nearest inner non-zero neighbours.) In leading and trailing sequences of zeros (where no proportion of the training data was found), the following modification occurred: at the innermost zero, the penalty was set to $-log(1E-10)$, and at each next-outermost zero the penalty was increased by $-log(1E-10)$.

In the transcription algorithm, whenever a path entered a phone, a count was begun of the number of frames spent in that phone, and on leaving the phone, the penalty was looked up for that number and added into the path-score (penalties were saved so they could be considered again in "remote search" (4.6.2) from beyond any TR class frames). Hence the probability of a phone's having some

particular duration (which is what I was attempting to approximate) was treated as independent of the probabilities of the frames concerned being frames from a phone of that class. (The precise relationship between these two probabilities — one applicable to individual frames, and one to a sequence of frames — is not entirely clear, and when, for example, one has a phone-hypothesis for a long vowel of 20 to 25 frames in duration, and one sets against the 20 or so probability-scores a single duration-based score, one has some misgivings as to whether one is actually giving an appropriate weight to the durational factor.)

The results of this experiment are shown in tables 6.7, 6.8 and 6.9. The results show that crude duration-modelling of this kind has a significant effect on the accuracy figure (approximately 3.3 points better with duration-penalties than without, with some 400 fewer insertions), but seems less than beneficial in terms of the percentage of phones correctly recognised (0.81% or 80 phones worse in the system with duration penalties), with nearly 100 more deletions. Naive duration-modelling of this kind appears to be useful in blocking implausibly short (and probably also implausibly long) phone-hypotheses, where the implausibility holds for *all* contexts of stress, clause-position, etc., but this amounts to a system of penalties for extreme duration-values rather than duration-modelling in any true sense of that term. (One may compare the following statement of Hervé Bourlard in this connection: "It seems that the best solution [to the duration-modelling problem in HMM] is also the simplest one and consists in imposing a (trained) minimum duration (simply by duplicating states that cannot be skipped)"(Bourlard 1995).) The principal work done by the penalties appears to be to block insertions of implausibly short segments, such as insertions of [ii] in velar pinched onsets of [e], insertions of [ax] and [r] in [r ax] sequences that formerly produced transcriptions with [r ax r ax], insertions of single-frame "right-sided" [w] in those contexts where right-sided [w] may occur, and substitutions of ai-glides for short schwas with dynamic F2-pattern.

It needs to be acknowledged that the implementation of the duration-penalty system could be further improved without a wholesale shift to something more sophisticated, though the extent of the improvement would be dependent on additional training-data being available. The present duration-penalties are, for one thing, almost everywhere insensitive to phonetic context and taking account

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subs's | ins's | total |
| 69.33 | 50.71 | 6956 | 388 | 2689 | 1868 | 10033 |

**Table 6.7.** No Duration Penalties

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.52 | 54.03 | 6875 | 485 | 2673 | 1454 | 10033 |

**Table 6.8.** With Duration Penalties Added

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.0 | 0.0 |
| Substitutions | 0.705389 | 0.589222 |
| Insertions | 1.0 | 0.0 |

**Table 6.9.** Confidence and Significance Measures

of such context would very probably lead to better results. Also, of course, dramatic improvement in the effectiveness of duration-penalties should come from explicit differentiation of pre-clause-boundary syllables. To illustrate the point, one may consider the example of releases of [+VOICE] stops, whose duration-statistics are distorted by inclusion of utterance-final cases. This, though, would be the beginning of a shift away from naive modelling to something that began to face up to the obvious fact that the purely phonetic realm is not autonomous with respect to the durations assigned to its elements. (The particular issue of pre-pausal and similar locations of utterances will be revisited briefly in section 6.9.)

## 6.5 Relative Duration Penalties — The Problem of "Gliditis"

### 6.5.1 The problem

The system manifested an extreme (one might say, pathological) predilection for seeing diphthongal glides of [ai], [ei] and [oi] where they were not. [ai] glides were substituted for offsets of [aa] and sometimes [uh] preceding consonants that typically cause a rising trend in F2, and sometimes for offsets of [a] similarly (as before velars); similarly [oi] glides were substituted for offsets of [oo] preceding consonants with this effect on F2, even where the effect can be very slight as before non-dark laterals; and similarly with [ei] glides, which were frequently substituted for offsets of [e] before consonants which typically cause a rising trend in F2. Substitutions of [ai] glides were also fairly common for schwas with rising F2, though some of these were preventable once duration-penalties were imposed, the shortest genuine [ai] glides in the training-data being considerably longer than the shortest genuine schwas. The introduction of crude absolute duration penalties also probably had some effect of reducing the degree of "gliditis", since penalties were imposed separately on the glide and on the 'D1' element, whereas monophthongal vowels had only a single duration-penalty imposed on them; the extent of any such effect could have been investigated by running the system

without absolute duration penalties but with the relative duration-penalties for diphthongs, but time prevented this.

### 6.5.2 Possible causes

A number of factors appeared likely to be responsible for this phenomenon of "gliditis". Firstly (cf 5.3.4 on bias), the offsets of the affected monophthongs tend to be rather weakly modelled in comparison with the diphthongal glides, since the former involve generalisation over a number of different consonantal contexts and over (typically) a number of vowels close to one another (see 5.3.1 and 5.3.2), while the diphthongal glides are defined in terms of both origin and a specific destination ([ii]-like, [i]-like, [e]-like or [ax]-like), and typically are modelled from large amounts of data (but see also 6.5.5 below). When glides follow a cognate 'D1' vowel, the other factor leading to easy confusibility is the frequent lack of any real distinction between the D1 and the corresponding monophthong pre-offset (e.g. between [aa] and [aaD1]). Confusibility with certain schwas is not particularly mysterious when duration is ignored, given schwa's 'transparency' to surrounding context (Bates 1995; Kondo 1995).

### 6.5.3 Possible solutions

Leaving to one side for the moment the problem of substitutions for schwa, once reflection had been forced on what actually distinguishes these diphthongs from the related monophthongs when no 'D3' target is sustained in the diphthong, it began to appear that perhaps the most crucial factor is actually the difference between (on the one hand) the relative duration of the non-glide and glide phases in the diphthongal case and (on the other hand) the relative duration of the pre-offset and offset phases in the monophthongal case; with diphthongs the ratio tends to be much smaller (glides being relatively long, while offsets of monophthongs are relatively short). One promising line of attack therefore appeared to be to take account of these relativities in calculating class-probabilities, and the most obvious strategy appeared to be to gather relative duration statistics and use them to modify path-scores during DP search, simply counting the numbers of frames in the subphones concerned in any local path and calculating a

"penalty" determined in accordance with the relative duration-statistics. There is an obvious problem with this, however: if such penalties are to be imposed on any pairs or triples of subphones, they ought to be imposed on all alike in order to maintain a level playing-field, and yet the possibility of getting reliable statistics for all cases was not open given the amount of data or indeed the time available.

Given that the problem was initially thought to be chiefly one of bias — of certain competitors being unfairly favoured — it was decided that (however imperfect the move might be in theoretical terms) it would still be worth seeing how far the problem could be righted by penalising *just* the diphthongs, or at least those without an hypothesised 'D3', on the basis of the relative durations of pre-glide and glide. The problem remained of what weight should be given to any penalties in such an ad hoc stratagem. There does not appear to be any "right answer" to this question: in each individual case (for each phone concerned) the degree of compensation that would be required to ensure a right answer will be different, and probably all that can be done in this approach is to determine by experiment what weighting gives the greatest reduction overall in the number of substitutions, which is quite unsatisfying from a theoretical point of view. A more satisfying approach might have been to *pool* the data for the confusible cases and likewise to pool the data for the pre-glide or pre-offset subphones and to let the relative duration be the sole arbiter, but again this would have required more data and time than was available.

### 6.5.4 The Experiments

The basic procedure for calculating duration-penalties was similar to that used for calculating absolute duration-penalties as described earlier in section 6.4. No distinction was made between diphthongs in gathering the relative duration statistics (one set of ratios was deemed applicable to [ai], [ei] and [oi]), but a distinction was made between glides to a schwa-like target and glides to other targets (it was thought likely, on the basis of visual inspection of the spectrographic data, that glides to schwa might tend to be shorter, relative to the pre-glide, than glides to

other targets). The numbers of cases falling into each of twenty bins were calculated, and these numbers were subsequently converted to relative frequencies. The lowest bin covered cases where the ratio (between duration of the D1 and the duration of the following glide) was less than 0.125, the second covered cases where the ratio was between 0.125 and 0.2, the third cases where the ratio was between 0.2 and 0.4, and then on in increments of 0.4 up to the fourteenth bin which covered cases where the ratio was between 4.0 and 4.4, after which the upper limits for succeeding bins increased each time by 1 (to 5.4, 6.4, and so on). Some smoothing then took place using a simple three-point moving average, but such smoothing was prevented from encroaching into leading or trailing strings of contiguous zeros (i.e. none of the zeros in such regions were smoothed). Once the bin-counts had been converted to relative frequencies, penalties were calculated as negative logs of the relative frequencies; in leading and trailing sequences of zeros (where no proportion of the training data was found), the following modification again occurred: at the innermost zero, the penalty was set to $-log(1E-10)$, and at each next-outermost zero the penalty was increased by $-log(1E-10)$. The penalties are displayed in tables 6.10 and 6.11.

Relative duration (RD) penalties were levied only in cases where a diphthong had no real 'D3' subphone (sustained target) and where the glide was not followed immediately by a non-cognate vowel. The motivation for the restriction was simply that the real competition faced by the affected monophthongs was from pre-consonantal diphthongs, so that there was no justification for burdening all 'three-phase' diphthongs everywhere with RD penalties.

Applying the penalties in the transcription algorithm is relatively straightforward. On entering a 'D1' an additional count is kept of the number of frames spent in the D1, and on leaving the D1 to enter a glide, the additional frame-count is carried forward, and on leaving the glide and taking count of the frames spent in the glide in turn, the ratio is taken and the penalty found by consulting the penalty for the appropriate relative frequency bin. (The penalties themselves are saved for the purposes of "remote search" (4.6.2 )). The penalties are added in to the path scores as if they were genuine (negative log) probabilities and we were treating the relative duration as an independent factor along with the absolute duration and the cepstral class-membership probability. All the RD penalties

| Ratio | Penalty |
|---|---|
| Less than 0.08 | 690 |
| 0.08 - 0.125 | 460 |
| 0.125 - 0.2 | 230 |
| 0.2 - 0.4 | 18 |
| 0.4 - 0.8 | 15 |
| 0.8 - 1.2 | 16 |
| 1.2 - 1.6 | 19 |
| 1.6 - 2.0 | 22 |
| 2.0 - 2.4 | 28 |
| 2.4 - 2.8 | 33 |
| 2.8 - 3.2 | 38 |
| 3.2 - 3.6 | 230 |
| 3.6 - 4.0 | 460 |
| 4.0 - 4.4 | 690 |
| 4.4 - 5.4 | 920 |
| 5.4 - 6.4 | 1150 |
| 6.4 - 7.4 | 1380 |
| 7.4 - 8.4 | 1610 |
| 8.4 - 9.4 | 1840 |
| 9.4 and above | 2070 |

**Table 6.10.** Ratios and Penalties for Diphthongs with glide to [ax]

| Ratio | Penalty |
|---|---|
| Less than 0.08 | 230 |
| 0.08 - 0.125 | 36 |
| 0.125 - 0.2 | 25 |
| 0.2 - 0.4 | 18 |
| 0.4 - 0.8 | 14 |
| 0.8 - 1.2 | 16 |
| 1.2 - 1.6 | 21 |
| 1.6 - 2.0 | 25 |
| 2.0 - 2.4 | 29 |
| 2.4 - 2.8 | 37 |
| 2.8 - 3.2 | 44 |
| 3.2 - 3.6 | 49 |
| 3.6 - 4.0 | 46 |
| 4.0 - 4.4 | 230 |
| 4.4 - 5.4 | 460 |
| 5.4 - 6.4 | 690 |
| 6.4 - 7.4 | 920 |
| 7.4 - 8.4 | 1150 |
| 8.4 - 9.4 | 1380 |
| 9.4 and above | 1610 |

**Table 6.11.** Ratios and Penalties for Diphthongs with glide to [ii], [i] or [e]

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.52 | 54.03 | 6875 | 485 | 2673 | 1454 | 10033 |

**Table 6.12.** Absolute Duration Penalties only

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.43 | 54.24 | 6866 | 485 | 2682 | 1424 | 10033 |

**Table 6.13.** With Relative Duration Penalties added

were weighted by a factor of 5 in this experiment. The results are shown in tables 6.12, 6.13 and 6.14.

The result is not very conclusive, with 10 fewer phones correctly recognised once relative duration penalties were introduced, but 30 fewer insertions (the result for insertions appears to be highly significant). The peculiar nature of the errors involved here needs to be remembered, however. These errors were of the type exemplified by misclassifications of [aa] as [aaD1 AIgl], and the 30 fewer insertions can probably be taken to be at least for the most part the results of [AIgl] and the like being blocked; on the other hand, the mere blocking of a path through [aaD1 AIgl] does not guarantee that the desired path through [aa] must come out best, since [aa] suffers from poor discriminability vis-a-vis [uh] and to a lesser extent [o], and also vis-a-vis [ax dl] when its left context is alveolar, palatal

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.5 | 1.0 |
| Substitutions | 0.23266 | 0.46532 |
| Insertions | 0.999456 | 0.001088 |

**Table 6.14.** Confidence and Significance Measures

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 67.65 | 53.86 | 6787 | 550 | 2696 | 1383 | 10033 |

**Table 6.15.** Absolute Duration Penalties only, 190 Training Sentences

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 67.63 | 53.89 | 6785 | 557 | 2691 | 1378 | 10033 |

**Table 6.16.** Relative Duration Penalties added, 190 Training Sentences

or velar. Hence the degree to which the particular problem of gliditis is addressed by the use of relative duration penalties is best judged by means of the figure for insertions. On the other hand, there were cases where diphthongs were *correctly* recognised prior to the introduction of relative duration penalties, and where the penalties caused them to be misrecognised.

Given that the problem was originally put down to the result of bias, and given that additional sentences had been incorporated in training from the outset to try to compensate for the bias (additional sentences being expressly chosen for the presence in them of affected monophthongs such as pre-alveolar [aa]), I decided to repeat the above experiment using only the 190 ATR training-utterances to see if the gap between the run with and the run without the relative duration penalties was greater with the smaller amount of training-data, a result which would lend strength to the hypothesis that the problem is indeed essentially one of bias. The results of this experiment are shown in tables 6.15 and 6.16.

Surprisingly (to me, at least), the relative duration penalties appear to be doing even less work when the more limited amount of training-data is used. The smaller amount of training-data (particularly of data for the affected monophthongs) ought, I thought, to have meant a greater degree of "gliditis" to be attacked by means of the relative duration penalties, but if it was present, the relative penalties seem to have been largely ineffective in dealing with it.

| Pairwise Closed-Test Discrimination Test Results | |
|---|---|
| Pair of Classes | Respective Scores (% correct) |
| Caa vs CaaD1 | 84 and 76 |
| Coo vs CooD1 | 85 and 85 |
| Ce vs CeD1 | 91 and 91 |
| AA_NVCORA vs AI_IIgl | 100 and 100 |
| AA_NVCORA vs AI_Igl | 98.5 and 98.7 |
| AA_NVCORA vs AI_Egl | 99.25 and 99.5 |
| AA_NVCORA vs AI_AXgl | 97.7 and 97.95 |
| AA_VCORA vs AI_IIgl | 100 and 100 |
| AA_VCORA vs AI_Igl | 96.6 and 93.65 |
| AA_VCORA vs AI_Egl | 96.67 and 97.82 |
| AA_VCORA vs AI_AXgl | 94.29 and 95.64 |

By increasing the multiplicative weight on the penalties, one could block additional substitutions of diphthongs, but even as it stands (with a weight of 5) and as already stated above, some true diphthongs are being misrecognised that were correctly recognised without the penalty. The use of relative duration penalties in this crude fashion, on a single class of phones, is clearly not a policy with a great deal to recommend it.

### 6.5.5 Frame-level Discriminability for the Phones Involved

I conducted a number of closed test pairwise classifications to see how much confusion there was between frames from the confusible classes when neither representativeness of data, nor bias, nor segmentation could be an issue. The results are shown in table 6.5.5.

The precise interpretation of these (closed-test) results may be uncertain, but one thing that does seem to be suggested by them is that it may be the discrimination between 'D1' and core monophthong, rather than that between glide and monophthongal offset, that is responsible for the major part of the problem (this conclusion is only immediately available for the [aa] vs [ai] case, as I did not test the other cases, but the extension to the other cases is plausible at least). Had time allowed, it would have been worth trying pooling the core 'D1'

| data-set | mean | standard deviation |
|---|---|---|
| A | 5.056 | 1.634 |
| B | 12.9 | 3.288 |
| C | 8.622 | 4.425 |

**Table 6.17.** Parameters for data-sets A, B and C = A+B

and core monophthong data and seeing whether this would reduce the degree of gliditis.

It is perhaps worth emphasising that the troublesome diphthongs (those without following 'D3' elements) represent a weak point in respect of sequencing-constraint. The glide is defined with respect to its target ([AI_IIgl], [AI_Igl], etc.), but not with respect to the identity of the following consonant or non-'D3' vowel, and this undoubtedly facilitates misrecognition of monophthongal offsets as glides.

## 6.6 Effects of the assumption of Normal distributions

The kind of damage that can result from inaccurate assumptions of Normality can be illustrated for the univariate case by means of figure 6.1, which plots normal curves for two sets of data A and B that have minimal overlap, together with a normal curve for the combined data-sets under the assumption that they belong to a single normal population. The parameters for the three data-sets are given in table 6.17.

The principle is no different with multivariate data, and the figure thus helps to illustrate one way in which misclassifications can arise where classes that are being assumed to be multivariate normal are actually not so. The figure illustrates a fairly extreme case (there are probably few cases in the data where so much separation will be found between modes of an assumed normal distribution as exists between data sets A and B in the example), but less extreme cases may still involve significant loss of accuracy.

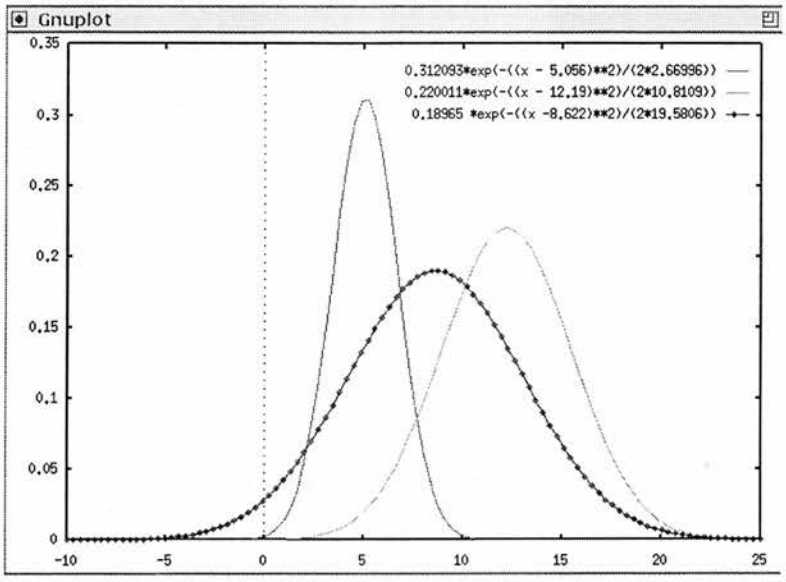



**Figure 6.1.** two normal populations treated as one

Some indication can be obtained of the likely cost to the system of working with single gaussian models by simple inspection of collections of cepstra produced by the training procedure for particular subphones. It is not, of course, true that normality can be accurately assessed by simple inspection,[6] but there are at least cases where it is clear from simple inspection that the data are very likely *not* normal. This is the case, for example, where modes are clearly evident for one or more variables, or where distributions for individual variables are skewed, or where the density for one or more variables appears to suggest something more akin to uniformity than to normality. It is perhaps worth noting here that the normality of each individual variable is a necessary condition for multivariate normality, though not a sufficient one; where a non-normality appears to be present for any individual variable, then, the normality of the joint distribution is thereby also thrown into doubt (while apparent normality for each individual variable does not guarantee multivariate normality). (Formal tests could also be applied to assess the normality of the data. Distributions for

[6]Once saturation is reached, for example, it is impossible to tell whether one region within the saturated area is more densely populated than another.

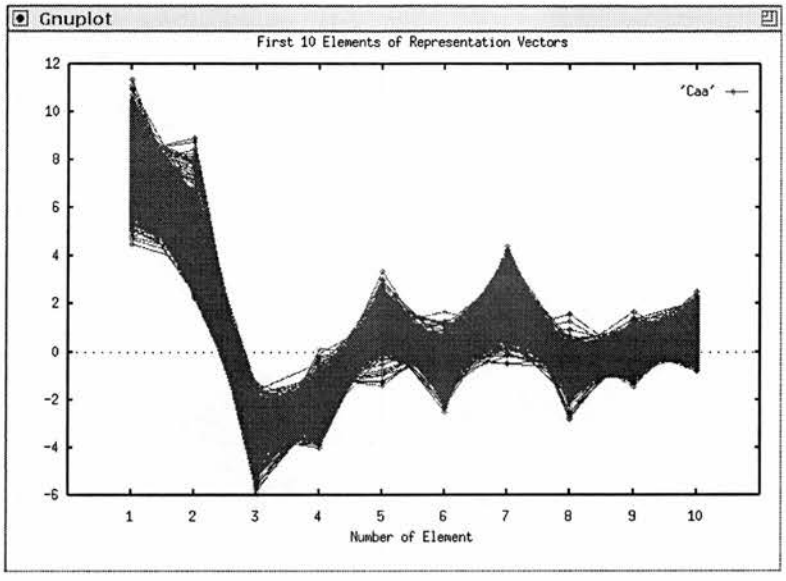



**Figure 6.2.** Static Cepstral Coefficients for [Caa]

individual variables could be tested using the Kolmogorov-Smirnov test, which measures the maximum distance between the sample and the theoretical cumulative probability functions. Measures of skewness for the multivariate normal distribution are given in (Mardia 1970).)

It was admitted in 3.6 that the assumption of normality for the delta coefficients is highly suspect, in many instances at least. Here the focus is rather on the static cepstral coefficients. Plots are provided of collections of cepstra for a number of classes. Cepstra for cores of three relatively stable vowels — [Cii], [Caa] and [Coo] – are given first in figures 6.2, 6.3 and 6.4 for orientation.

For these three cases, the normality assumption does not appear to be grossly unjustified, as far as it is possible to tell. In the case of a small number of vowels, however, (even ignoring the extremely variable schwa), the normality assumption can be seen to be less than ideal even for the static coefficients for the core subphones. It is worth recalling to begin with that the decision to create a single core state for each vowel, regardless of its left and right context, was prompted largely by the insufficiency of data for an attempt at anything more sophisticated, rather than from any belief that whatever steady-state may be reached will be

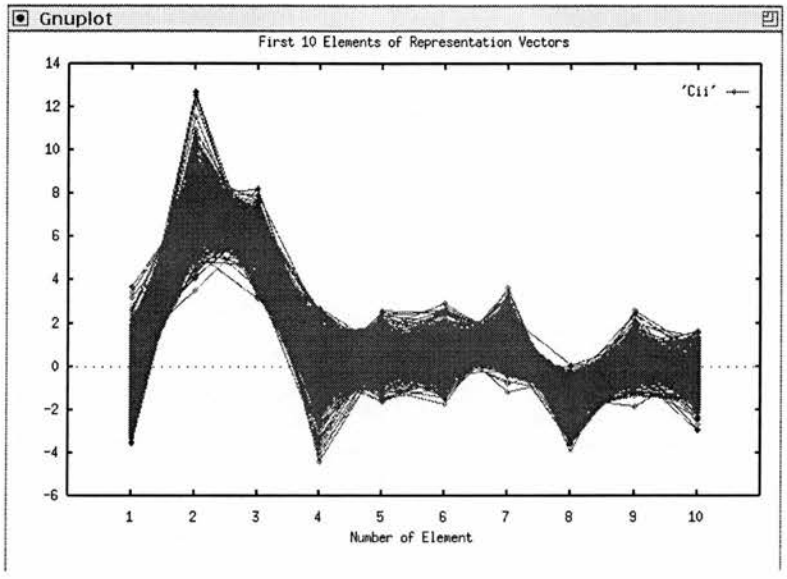



**Figure 6.3.** Static Cepstral Coefficients for [Cii]

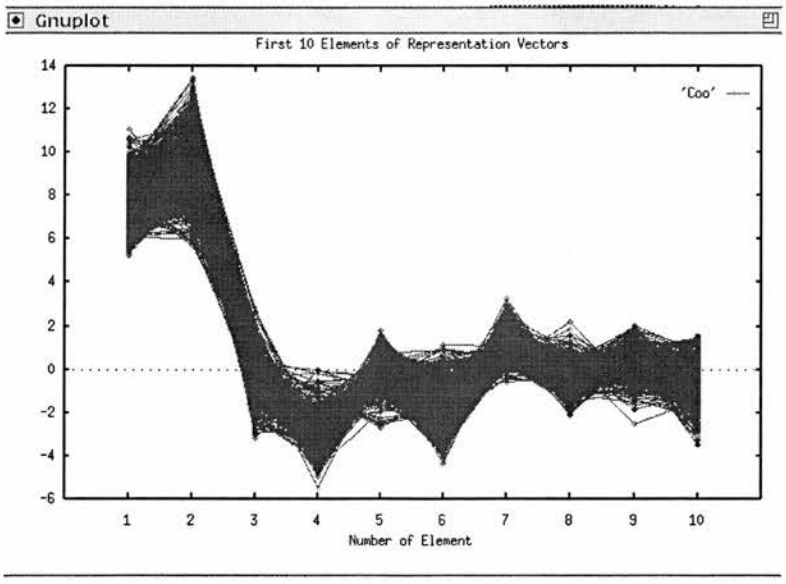



**Figure 6.4.** Static Cepstral Coefficients for [Coo]

independent of the preceding and following context. In a number of cases, though, the effects of preceding context, or in the case of /uu/ an extraordinarily wide range of possible realisations, prompted me to create distinguishing labels that allowed separate modelling of these cases, thus helping to preserve to some extent the reasonableness of the normality assumption. Thus [ii] immediately following [r] was labelled as [rii], and the /uu/ vowel was labelled as [uuf], [uum] or [uu] depending on the height of F2 (cf 2.5.9). (A further case that ought to have been dealt with similarly was that of [i] preceding [ng]; in a majority of cases [i] in this context has a very [ii]-like realisation as far as its spectral pattern is concerned, and is consistently misclassified as [ii] accordingly in such cases — one could of course have blocked [ii ng] strings on phonotactic grounds, but was loath to do so given the possibility of "phonetic" [ii ng] in productions of words like "seeing".) When cepstra for [Crii], [Cuuf], [Cuum], and [Cuu] are plotted, the weakness of these sorts of ad hoc props for the Normality assumption becomes apparent. The "rii" label was used automatically in manual transcription of the training set whenever an [ii] was preceded by an [r], but in reality the extent to which the F-pattern of [r] comes to influence or dominate that of following [ii] varies from case to case. In the case of the "allophones" of /uu/, the cut-off points for labelling were not precise, and it is fairly clear from the plots that there is some overlap between the three. The problems are illustrated in figures 6.5 to 6.12, the first three representing the 'artificially' separated forms of /uu/.

([Cuu] is clearly very under-trained, and it is conceivable that its non-normality would become less severe as training-data was increased.) The plot for [Crii] suggests a minority of vectors which are distinctive, at least in respect of the 4th coefficient. The reason the case is no more extreme than it is is probably due to the fact that in training vectors which are some distance from the norm for the core state may be allocated to onset or offset, with core subphones being optional in the event of an onset being present. Similar considerations would appear to be relevant to the data for [Ci], whose plot (figure 6.9)does not suggest any striking non-normality for any coefficient. Plots of associated subphones [R_RIIB][7] and [HF2_VELA][8] (which covers [i_ngA] as a particular case) are therefore shown in

[7]onset of [rii] or [riiD1] following an [r] or 'xrb'

[8]offset of [i] or [iD3] or [ir] or [irD3] or [uuf] before a velar consonant

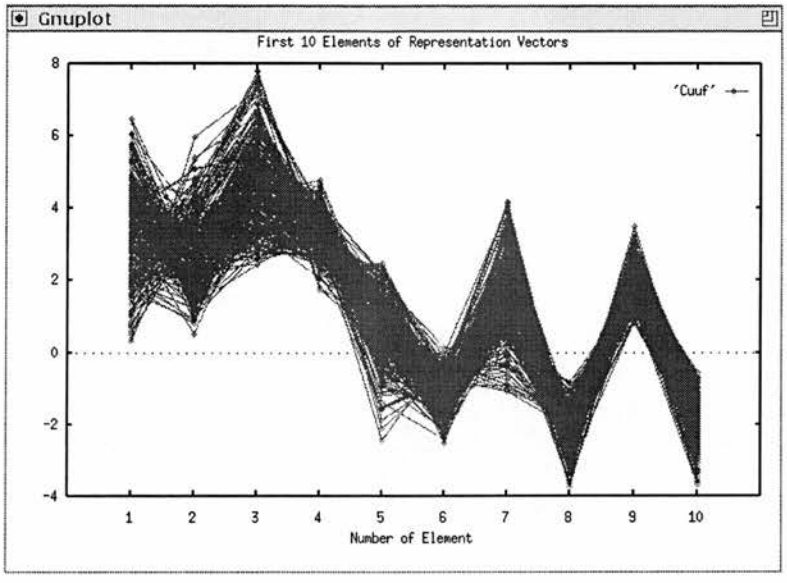



**Figure 6.5.** Static Cepstral Coefficients for [Cuuf]

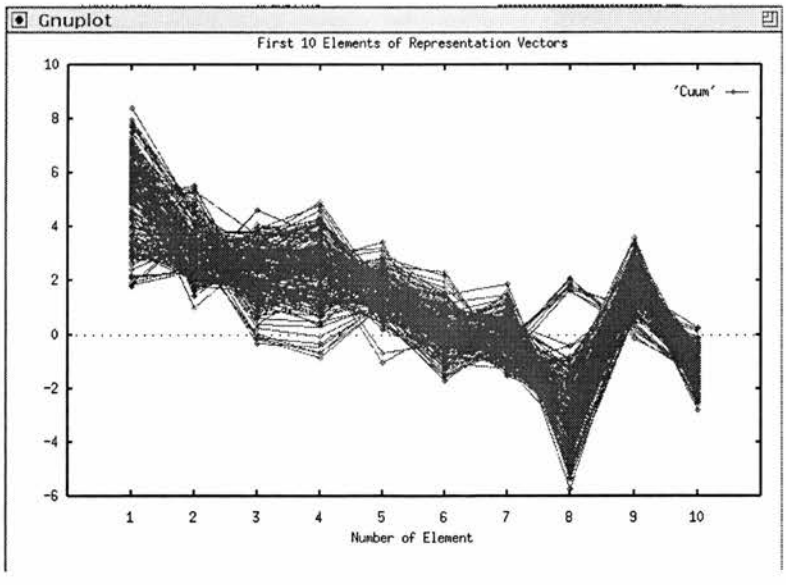



**Figure 6.6.** Static Cepstral Coefficients for [Cuum]

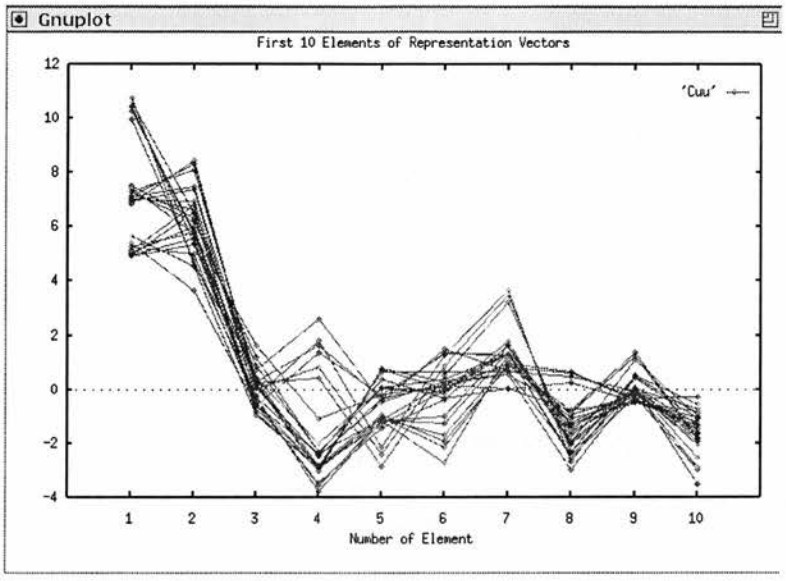



**Figure 6.7.** Static Cepstral Coefficients for [Cuu]

addition. The plot for [R_RIIB] (figure 6.12) suggests skewness to the left (towards lower values) for the first and perhaps the fifth coefficients, and skewness to the right for the seventh. The plot for [HF2_VELA] (figure 6.10) shows skewness to the left for the first coefficient, and suggests a degree of skewness to the right for the third and seventh coefficients, with an apparent dip in density about the middle of the range for the fourth coefficient; there also appears to be a minor outlying mode in the distribution for the eighth coefficient. How much the apparent non-normality of [HF2_VELA] arises from the putting together of a nasal and non-nasal right contexts, how much from the putting together of the four vowels [i], [ir], [rii] and [uuf], and how much from the tendency of [i] to become [ii]-like before [ng], is a question which there was not time to investigate.

The variable extent to which [r] may dominate the spectral pattern of a following vowel is perhaps one of the factors responsible for the non-normalities of many of the individual coefficients in the plot for [R_IB][9] (figure 6.11).

It may be of some interest also, in view of the 'gliditis' problem discussed

[9]onset of [i] following an [r] or 'xrb'

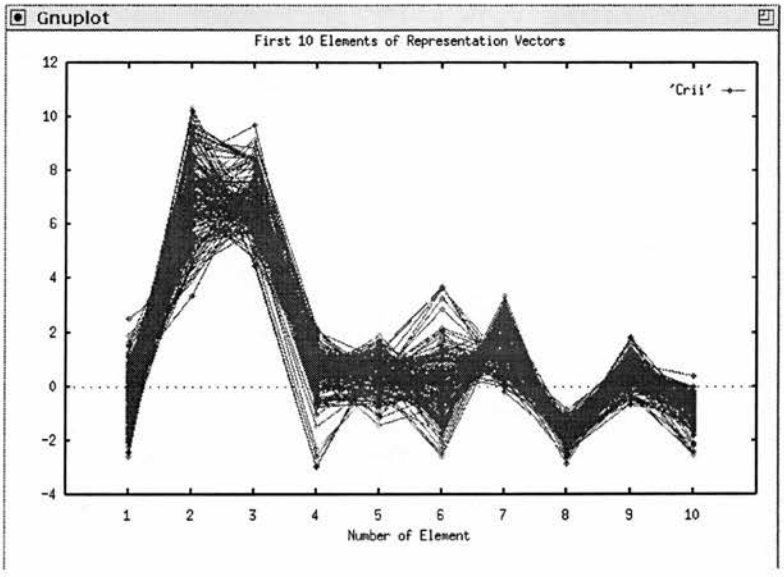



**Figure 6.8.** Static Cepstral Coefficients for [Crii]

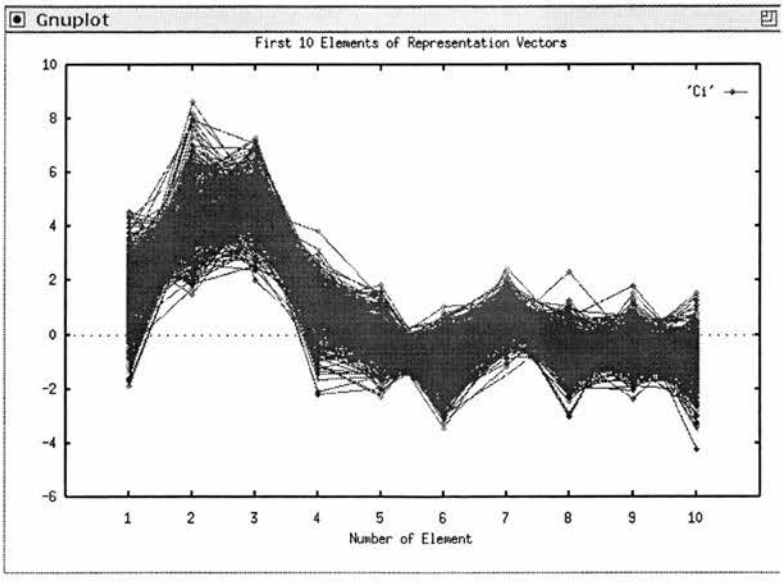



**Figure 6.9.** Static Cepstral Coefficients for [Ci]

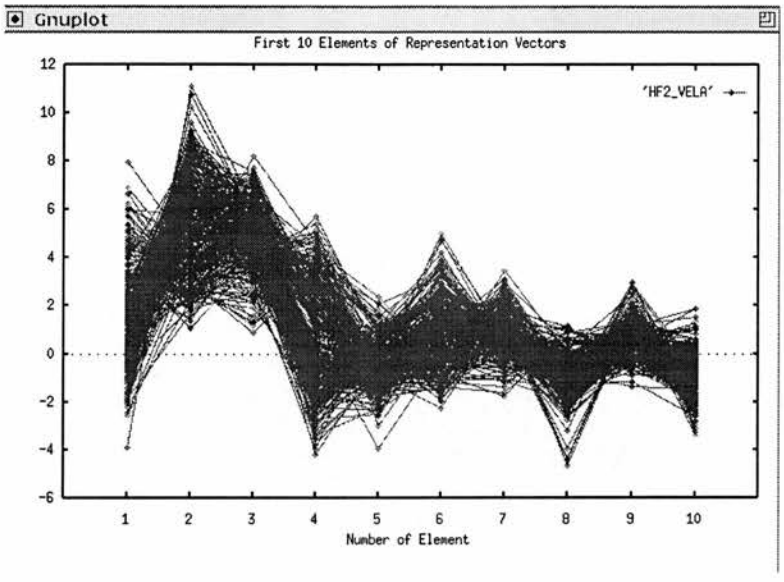



**Figure 6.10.** Static Cepstral Coefficients for [HF2_VELA]

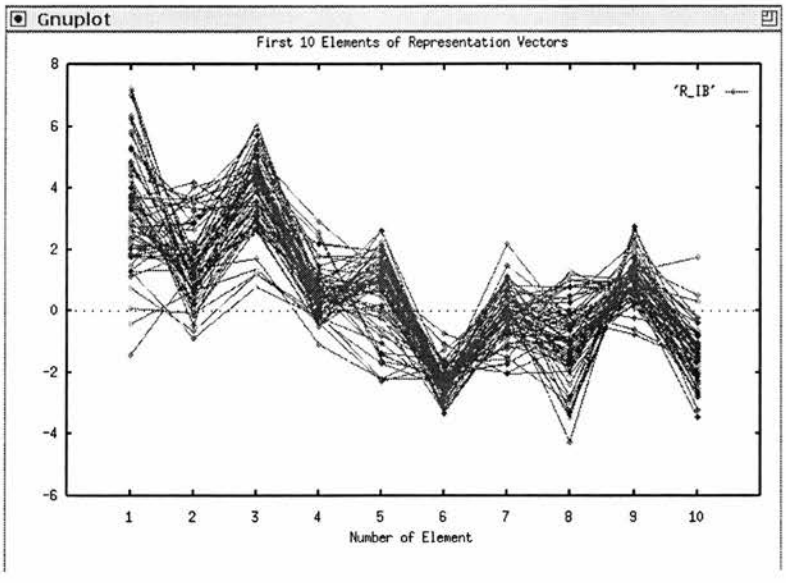



**Figure 6.11.** Static Cepstral Coefficients for [R_IB]

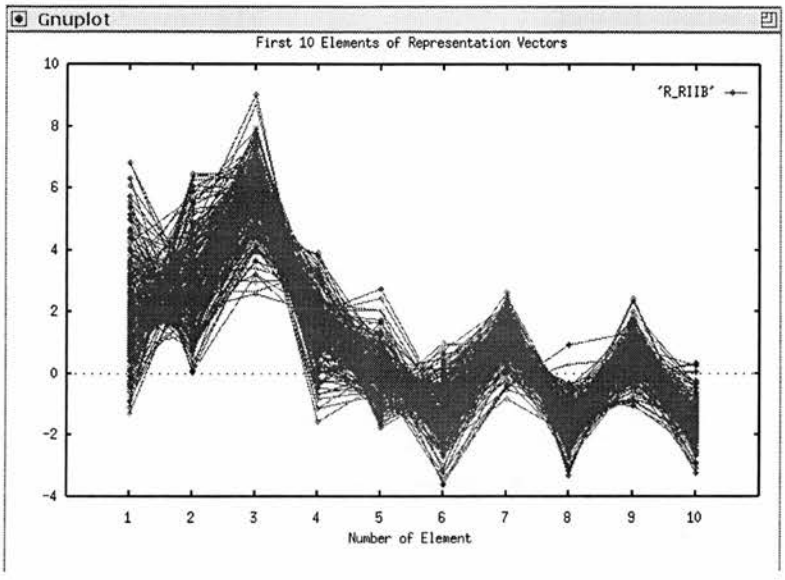



**Figure 6.12.** Static Cepstral Coefficients for [R_RIIB]

earlier, to look at a plot for [AA_NVCORA],[10] one of the classes subject to misclassification as a glide, shown as figure 6.13. The 7th coefficient comes close to the condition used to illustrate the danger of the normality assumption at the beginning of this section, with a clear dip in the density between two modes (the distributions for each of the first two coefficients also look markedly asymmetric). The plot for [AA_VCORA], given as figure 6.14, also suggests non-normality for a number of individual coefficients. The second, third and fourth appear skewed to the left while the seventh appears to be skewed to the right.

All the vowel-offsets with [VCORA] as their right context combine nasal and non-nasal contexts, and this may well be a major factor in their distributions not being normal. The next plot in the series is for another class in this category, [HB2_VCORA][11] (figure 6.15). It is possible to discern a number of modes in its plot, most notably at the third coefficient.

Finally, plots are given for a number of classes chosen from amongst two

[10]offset of [aa] before a [-VOICE] alveolar or palato-alveolar consonant

[11]offset of [uu] or [uum] or [uumD2] or [axrD2] before a [+VOICE] alveolar or palato-alveolar consonant

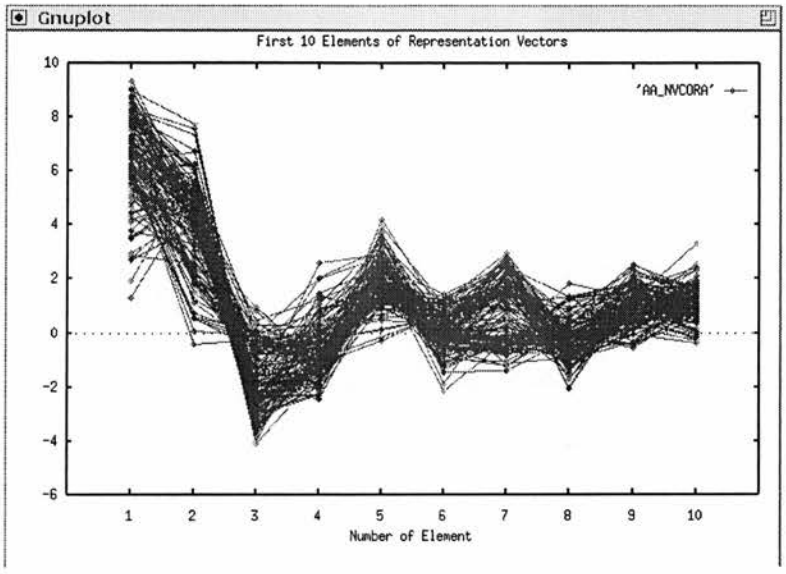



**Figure 6.13.** Static Cepstral Coefficients for [AA_NVCORA]

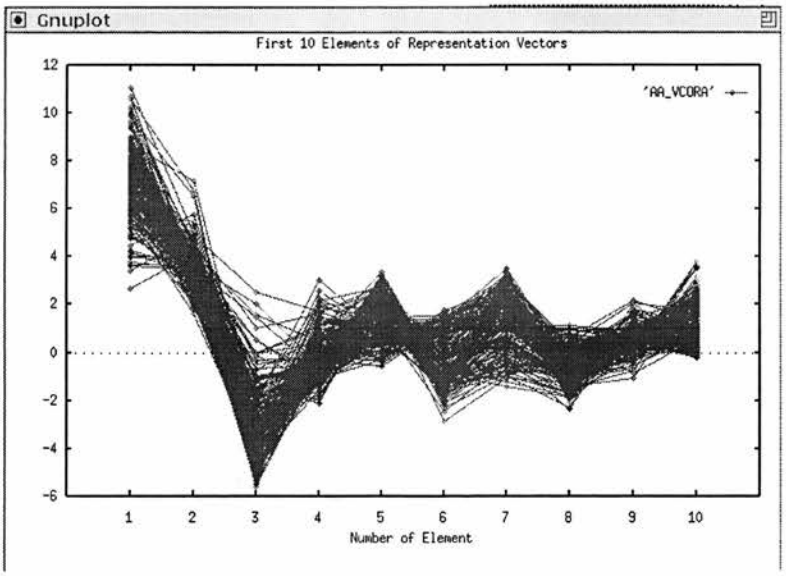



**Figure 6.14.** Static Cepstral Coefficients for [AA_VCORA]

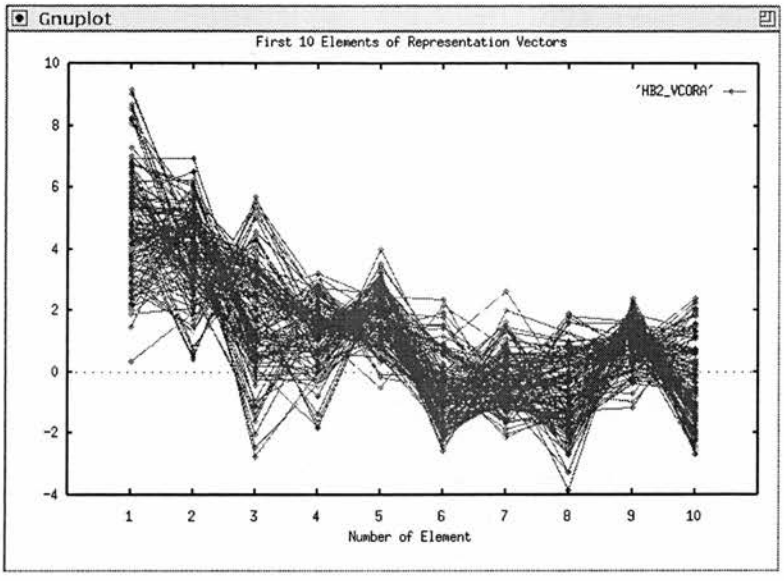



**Figure 6.15.** Static Cepstral Coefficients for [HB2_VCORA]

groups of classes thought likely to be liable to non-normalities, namely second phase subphones of prevocalic [-VOICE] stop-releases, and TR classes not resolved into 'A', 'B' and 'C' forms. [-VOICE] stop-releases, particularly for [kb], are likely to show late transients in spectrographic data, and these represent problems for any attempt to model the whole of the late part of the stop-release with a single gaussian. This is in addition, of course, to the generally rather unpredictable character of the spectral detail of these subphones. For some of the coefficients it is not absolutely clear in either plot that the densities are actually greater at any one point than at any other within the ranges covered; for the third coefficient of [kb2_MHBA][12] meanwhile, the distribution is clearly skewed to the left.

[TR10] models the border between nasals on the left hand and any of [r], [w], laterals or nasals on the right, and hence seems a good candidate for non-normality. Its plot is shown as figure 6.18.

[TR60] models the border between any pair of vowels, and was introduced at a

[12]the second phase of the release of a [k] preceding an [oo], [ooD1], [u], [uD1], [uu], or [uum]

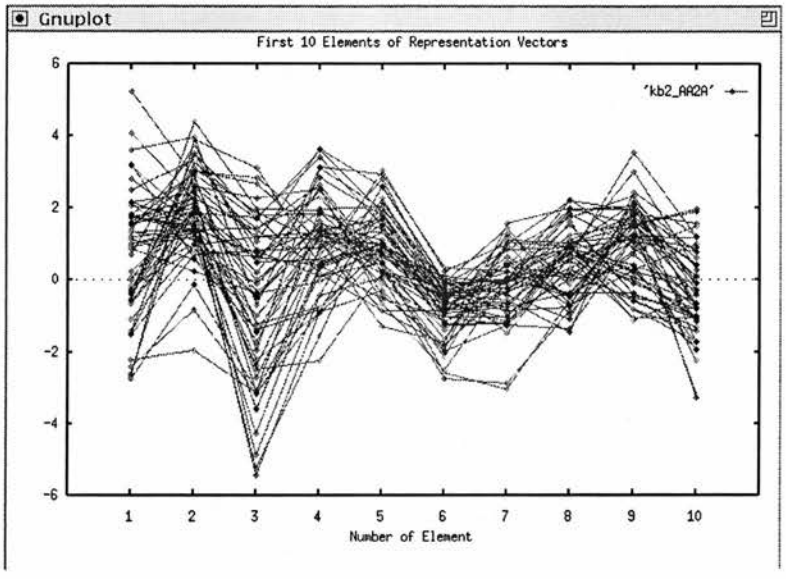



**Figure 6.16.** Static Cepstral Coefficients for [kb2_AA2A]

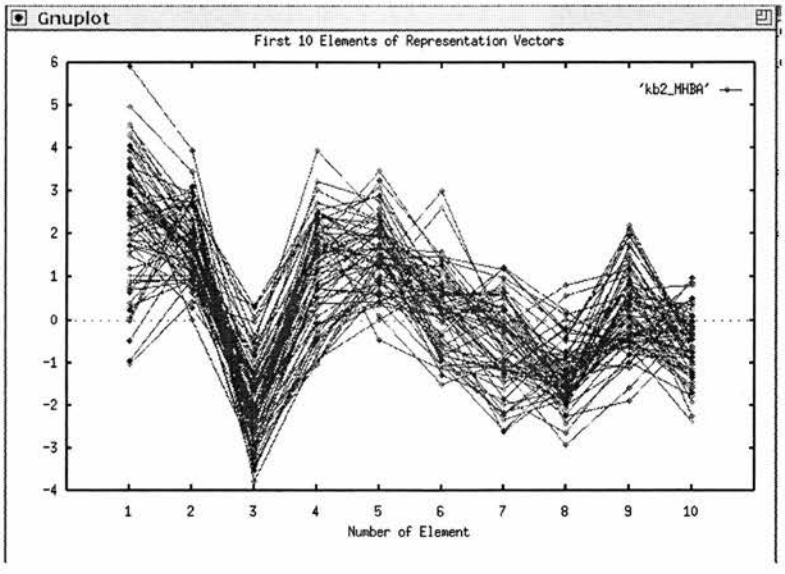



**Figure 6.17.** Static Cepstral Coefficients for [kb2_MHBA]

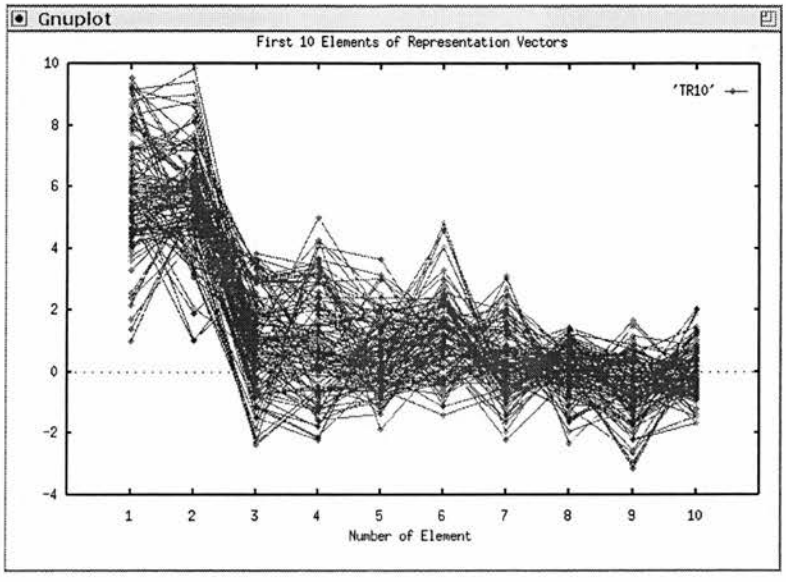



**Figure 6.18.** Static Cepstral Coefficients for [TR10]

late stage in an attempt to make insertions of spurious micro-vowels less likely; it is therefore far from representing a homogeneous cluster of acoustic effects (apart from bare vocality), and non-normality is again to be expected (figure 6.19). The first coefficient's distribution appears to be skewed to the left (the second coefficient's perhaps likewise), while there appear to be modes in the distribution for the third coefficient.

In some cases it is possible that the existence of modes in the data is only a function of under-training, and that the modes would disappear with the addition of further data, but where generalisation has occurred across several contexts, or where — as after palatals or velars sometimes in the onsets of vowels — a brief spectral pattern maintains itself before the major transition to the vowel's target-value, the existence of modes, or at least of asymmetric distributions, is to be expected, and calls either for mixture-modelling, the provision of additional subphonic elements, or perhaps an algorithm which "grows" subphones as required to get good modelling throughout subject to the normality assumption.

Quantification of the cost to the system of the normality assumption is difficult. One could approach the problem by incorporating mixture-modelling in the

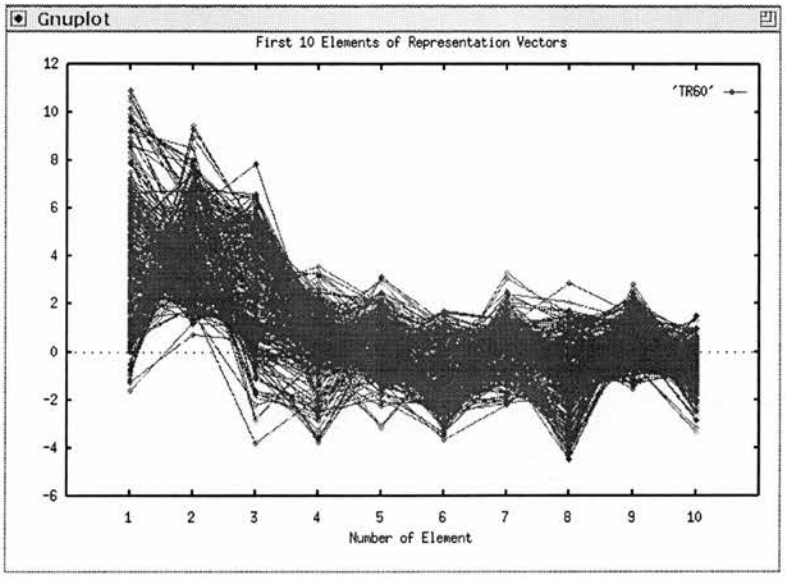



**Figure 6.19.** Static Cepstral Coefficients for [TR60]

system as it stands at present, and seeing how much improvement accrued, but this would require significantly more data in order to be feasible. The normality assumption certainly is sub-optimal (even if it is not clear *how much* blame it shares for the system's performance), and there is no obvious reason why the technique described here could not be used in conjunction with mixture-modelling, or with some other refinement designed to achieve the same end, with significant improvements likely as a result.

## 6.7 The handling of "TR" classes, Part II

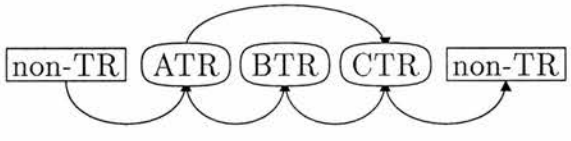
The nominal resolution of TR classes into separate elements described in 6.2 led to a significant degree of improvement in the proportion of phones correctly recognised. In this section I describe experiments designed to test whether further improvement could be gained by resolving TR classes not only nominally but spectrally too, modelling each of the two or three elements with the help of distinct data, at least where the training-data for the class in question was sufficient to allow this to be done. Two different experiments were tried.

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.43 | 54.24 | 6866 | 485 | 2682 | 1424 | 10033 |

**Table 6.18.** With Relative Duration Penalties, Nominal Resolution of TR's

In the first experiment, the allocation of frames to TR elements was as follows: in the case of a transition having three frames allocated to it, the first frame was assigned to the appropriate ATR class, the second to the BTR class, and the third to the CTR class; in cases involving only two transitional frames, the first frame was allocated to the appropriate ATR and the second to the appropriate CTR class. When all the training-data had been initialised in this way, a review was effected to see whether data-counts for these ATR, BTR and CTR classes were sufficient to make them viable, the criterion being that the BTR element should have at least 80 vectors assigned to it (because of the highly generalised nature of the TR classes, it was thought necessary to have a relatively high quorum-figure). For any TR class that failed to meet this criterion, merely nominal resolution was fallen back upon, the data for the ATR, BTR and CTR elements being pooled. As it turned out, only 14 of the 50 or so TR classes met the criterion for individual modelling of elements, which perhaps would predispose one to think that the results of the experiment would not be dramatic.

Sequencing rules were altered for this and for the later experiment (and applied regardless of whether a TR class was really or merely nominally resolved into discrete elements, though where the resolution was merely nominal the changed sequence-rules were equivalent to the old ones in practice). The changed sequence-rules can be illustrated by the following diagram:



| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.81 | 54.94 | 6904 | 472 | 2657 | 1392 | 10033 |

**Table 6.19.** With Relative Duration Penalties, Real Resolution of TR's (method 1(a))

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.941083 | 0.117834 |
| Substitutions | 0.922933 | 0.154134 |
| Insertions | 0.987659 | 0.024682 |

**Table 6.20.** Confidence and Significance Measures

This arrangement better reflects the way the individual TR elements are built up from training-data. The results of this experiment are given in tables 6.18, 6.19 and 6.20. A modest improvement in absolute terms is apparent once some concrete (as opposed to merely nominal) resolution of TR classes has been implemented, with 32 fewer insertions, 25 fewer substitutions, and 13 fewer deletions, though only the insertions result comes close to being significant at the level chosen. It still seems reasonable to conclude that with more data, which would allow more of the TR classes to be modelled in this way, the improvement might well turn out to be significant.

I thought it was worth trying method 1 again with a more courageous (and possibly foolhardy) threshold for going ahead with real resolution, and the result of trying again with a threshold of just 45 vectors is shown in tables 6.21 and 6.22. The number of classes really resolved using this threshold was typically 25. (The c-value and P-value refer to the comparison with the control, not to a comparison with the the test using a higher threshold.)

This threshold appears to fare a little worse in respect of correctness, with 18 more substitutions, but a little better in respect of accuracy, with 19 fewer insertions than with the higher threshold (and 51 fewer insertions than in the control

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.66 | 54.98 | 6889 | 469 | 2675 | 1373 | 10033 |

**Table 6.21.** Real Resolution of TR's, method 1(b)

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.856692 | 0.286616 |
| Substitutions | 0.684954 | 0.630092 |
| Insertions | 0.999745 | 0.00051 |

**Table 6.22.** Confidence and Significance Measures

experiment with merely nominal resolution of TR classes). The improvement in the number of insertions from the control to this condition is highly significant. An optimal threshold figure could have been identified, but there was not time to pursue the matter further.

In the second experiment a slightly different approach was tried, motivated largely by misgivings arising from the highly generalised nature of the TR classes and the suspicion that too much weight should not be put upon them, particularly where quotas were modest. The second experiment differed from the first chiefly in the initial allocation of frames to classes. Here, the procedure for three-frame transitions was as follows: the first and second frames were allocated to the appropriate ATR element, and the second and third frames to the appropriate CTR element, while all three frames were also assigned to the BTR element. For two-frame transitions, the same procedure was used as in the first experiment (no frames going to a BTR class). The criterion for going ahead with specific modelling was that the ATR element should have at least 100 vectors, failing which merely nominal resolution was fallen back upon. The number of TR classes that remained resolved spectrally was typically 28 under this scheme. The sequence-constraints, as mentioned above, remained the same as in the first experiment.

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.60 | 55.0 | 6883 | 476 | 2674 | 1365 | 10033 |

**Table 6.23.** Real Resolution of TR's, method 2

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.702872 | 0.594256 |
| Substitutions | 0.869886 | 0.260228 |
| Insertions | 0.975965 | 0.04807 |

**Table 6.24.** Confidence and Significance Measures

The results for this experiment are shown in table 6.23 and confidence and significance figures (for improvement from method 1 with the higher threshold) are given in 6.24. If we take the absolute figures at face value, the second method appears to be a little less good in respect of the percent correct figure, with 17 more substitutions than occurred with the first method, but a little better in respect of accuracy, with nearly 30 fewer insertions. Only the figure for insertions comes anywhere near to being significant, however. It is not possible, without further investigations, to draw conclusions about reasons for the detailed results from this rather slender basis, except perhaps to say that the matter could be investigated further. It was decided (a little arbitrarily, perhaps) to proceed using Method 1 with the higher threshold as the control for the next experiment.

An avenue that is surely worth investigating (and which was left unexplored only for lack of time) is that of making the TR's themselves less generalised. As an example, the highly generalised [TR1] class for borders between vowels and vowel-like phones on the left hand and stop-closures on the right could be differentiated even with existing amounts of training-data to take account of the PLACE of the stop, and this might well be found to contribute to reduction in the error in stop-identification following [ii], [iiD3], [uuf], [uum], and [uumD2] where there is little or no offset to provide cues, with what little offset there may

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 88.27 | 83.98 | 9015 | 264 | 934 | 438 | 10033 |

**Table 6.25.** Closed Test Results

be being swallowed by the non-specific TR class, leading to numerous errors at this point (cf 2.5.9, paragraph 6). With really massive amounts of training-data, of course, there would be a case for making all TR classes everywhere specific as far as possible, and it seems reasonable to expect significant benefit from doing so.

## 6.8 Representativeness of Training Data

Any evaluation of the system ought to involve consideration of the question of how representative the training-data is. Ideally, one would like to be able to provide some indication of what share of the system- error may be attributable to under-representation, and to arrive at reasonable projections of how performance might be expected to improve with increases in the amount of training-data, but these are difficult problems. It may perhaps be possible to argue that a closed test result could be taken as an indication of the best performance that could be achieved by the system given *unlimited* amounts of training-data, and thus (assuming a closed test performance figure of less than 100%) to put a limit to the room there may be for claiming that if only one had much more data, all one's problems would be solved. Closed test results for the 200 ATR sentences plus 1 TIMIT sentence are shown in table 6.25.

As would be expected, there is a dramatic difference between the closed test result and any of the open test results (though perhaps what is more interesting about the results is how low the headline figures are, rather than how high they are). In the closed test, the training-data is "perfectly representative", thus evading perhaps the most basic difficulty of all in phonetic classification, the variability between individual tokens of the same phonetic type. In open test, we

proceed on the assumption that our training-tokens are sufficient to cover the full range of such variability, and moreover to give us good estimates of the way the various possible forms are distributed, but of course these assumptions depend critically on our having training-samples that are adequate to the task, and we have seen reason for thinking that we almost certainly do not have such samples in every case (5.2). However, the closed-test result is certainly a pointer to the fact that merely increasing the training-data substantially would not result in all the problems disappearing.

The problem of poorly representative data is still likely to be a real one. Identifying cases where it is the major cause of transcription-errors is time-consuming, however, and it is not always possible to pin the blame squarely upon poor representation by the training-data, since other factors such as the single gaussian modelling assumption may also be relevant. In fact, of course, it is only on the assumption of representativeness that we are able to say of any data that appears non-normal that its population is non-normal, and when the amount of sample data is very small there must be doubts about the validity of such an assumption. Hence, given a plot for a limited sample that manifests (for example) pronounced modes for one or more coefficients, we cannot know *a priori* whether a doubling (say) in the amount of training-data would merely consolidate the modes, or on the contrary fill out the gaps between them to produce a more normal-looking distribution. There is an uncertainty regarding what sample-size constitutes a reasonable basis for parameter-estimation in the present context, just because of the many-to-one relationship between speech-frames and phonetic events (a point touched on early in chapter 5). Clearly a point must be reached where it is reasonable to assume that one's sample covers the full range of possible forms for the data, but this point may come quite late in the present context.

Two cases where it was considered possible that underepresentation might be at the heart of the problem are provided to illustrate these points, both involving particular (token) subphones that were consistently misrecognised (leading typically to error across a wider span than the subphone concerned, given the DP-based approach to utterance-transcription – it should be recalled that even a single very badly scoring vector may be sufficient to cause the best-scoring path to pass elsewhere than through the correct subphone).

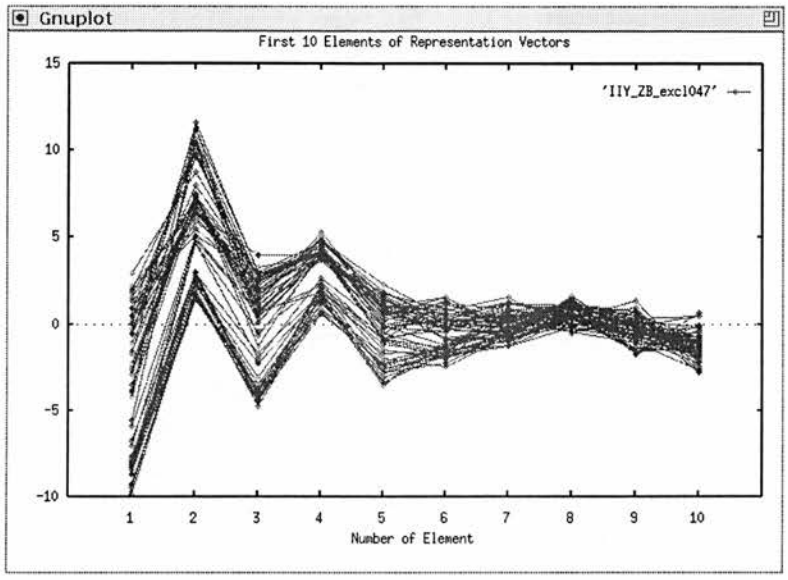



**Figure 6.20.** Training Data (static coefficients) for [IIY_zB]

The first of these cases was the onset of a [z] in 'these' in utterance 047. The frame-level scores given to the correct class ([IIY_zB]) were consistently extremely bad before the introduction of the additional 60 training-sentences (typically above 60 or 70, where the negative log of $1E-20$ is little more than 46), and the subphone was still consistently misrecognised even after the inclusion of the additional data. The most likely reason for this seemed to me to be — given the variability of post-vocalic onsets of [z] — that the training-data just did not happen to contain any examples that were similar to the token in question (the two main dimensions of variability in this case are persistence of voicing into the [z], and persistence of formant-structure). However, after isolating and plotting just the static coefficients for the training-data and test-item data, it appeared uncertain whether poor representation or non-normality was chiefly to blame. The plots are given as figures 6.20 and 6.21, and attention is drawn to the pronounced modes in the training-data plot, particularly for the first 4 coefficients. The actual values of the test-token's coefficients do not appear to lie outside the ranges covered by the training-data (though the 10th coefficient for one vector lies at the very fringe of the training-data distribution).

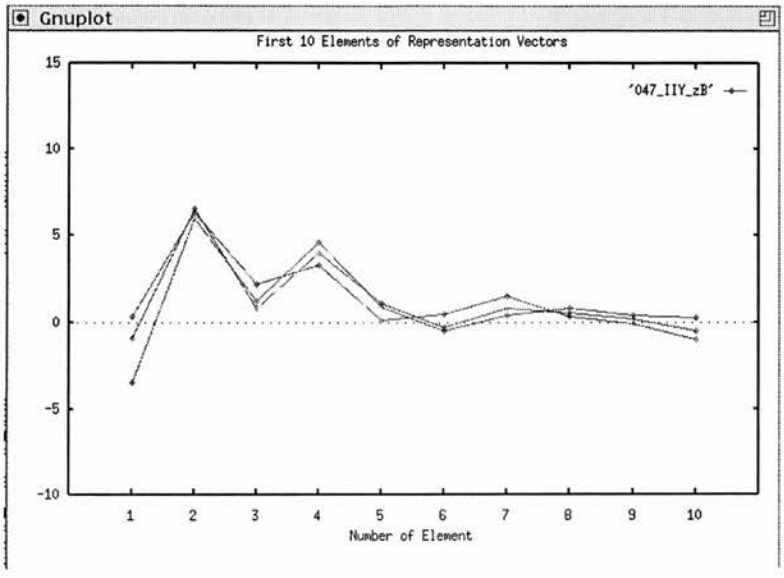



**Figure 6.21.** Test token (static coefficients) [IIY_zB]

Plotting the dynamic coefficients, however, made it look as if the source of the trouble did lie in under-representation. The plots are given as figure 6.22 and figure 6.23, and a combined plot is given as figure 6.24. Attention is drawn to the values for the first dynamic coefficient for two of the three test-vectors, which clearly lie outside the range covered by the training-data.

The second example considered is that of an offset of [e] before [th] in the word 'Beth' at the beginning of utterance 099. [th] was clustered with alveolar [-VOICE] consonants as a right context of vowels, a merge forced by data-shortage rather than chosen with enthusiasm. Although the word 'Beth' occurs right at the beginning on this utterance, the vowel-offset manifests the extreme devoicing associated more often with offsets of vowels before [-VOICE] consonants at clause-boundary or sentence-final position. In almost every one of the experiments above, this [E_NVCORA] subphone was recognised as [h ax] (giving [b e h ax th]).

The plot of the static coefficients for training-data for this subphone is given as figure 6.25, and the plot of the static coefficients for all the cepstra for the test-token as figure 6.26. Attention is drawn to the sixth coefficient on each plot: some

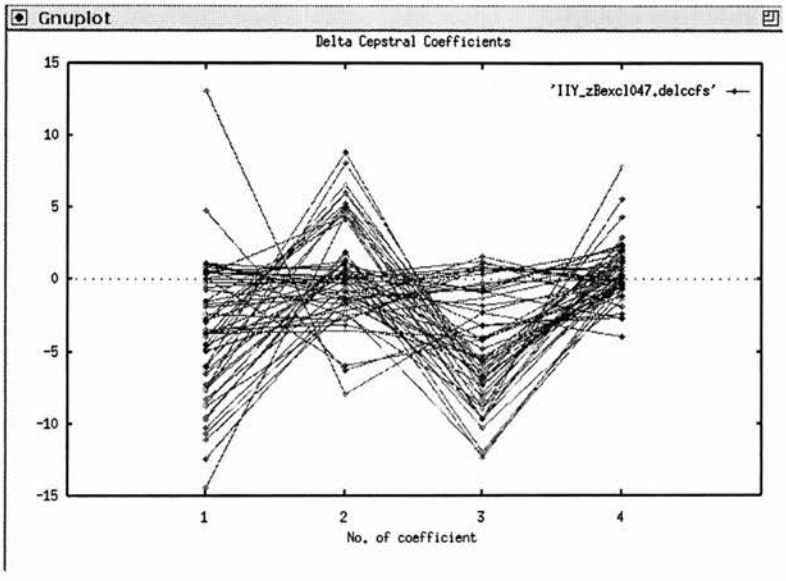



**Figure 6.22.** Training Data (dynamic coefficients) for [IIY_zB]

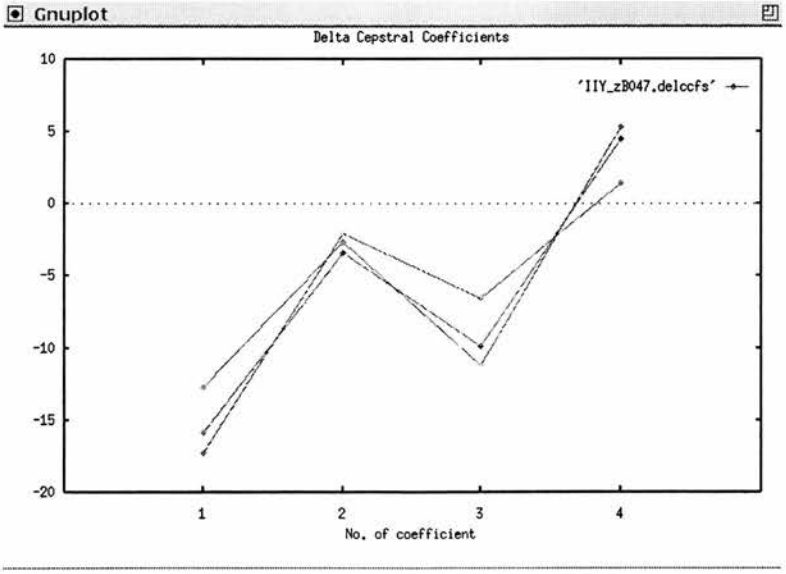



**Figure 6.23.** Test token (dynamic coefficients) [IIY_zB]

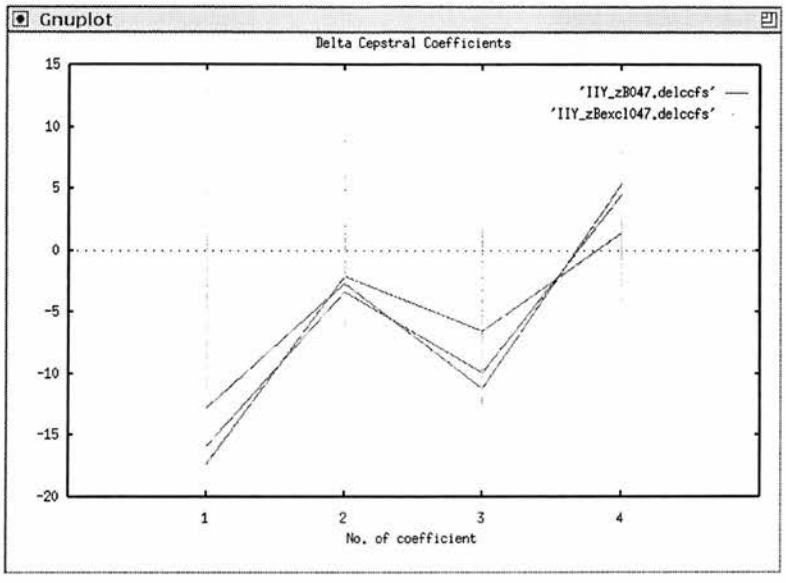



**Figure 6.24.** Combined plot of Training and Test token data (dynamic coefficients) for [IIY_zB]

third of the test-token vectors have values which lie in a tail of the distribution for the training-data, and it may be that this alone would be sufficient to destroy the chances of [e_NVCORA] beating off all competitors. One or two of the values for the 9th coefficient also look as if they might share some of the blame.

Plots for the dynamic coefficients are given as figures 6.27, 6.28 and (the combined plots) 6.29. Attention is drawn to two values of the first dynamic coefficient in the test-token data which lie outside the range of the values of this coefficient covered by the training-data. Once again it seems as if one of the dynamic coefficients shares most of the blame for the low probability score.

It has already been acknowledged in this work (2.5.9) that vowel-offsets before [-VOICE] phones may present difficulties for the three-subphone analysis, at least when modelling with single gaussians, and the plot for the test-token in this instance is testimony to the fact of a qualitative change during the temporal evolution of this offset.

Before leaving this section, it is perhaps worth noting the small increase in performance accruing from the increase in training-data of 60 sentences. In the

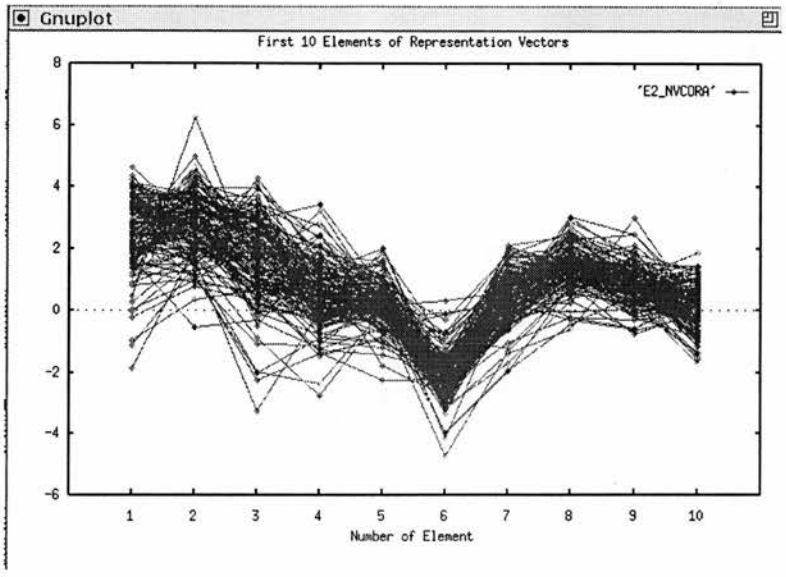



**Figure 6.25.** Training Data (static coefficients) for [E2_NVCORA]

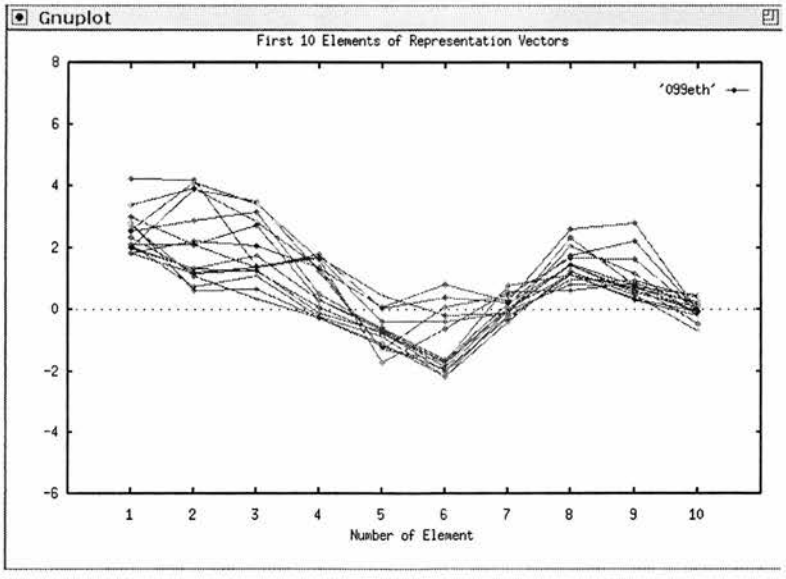



**Figure 6.26.** Test token (static coefficients) [E2_NVCORA]

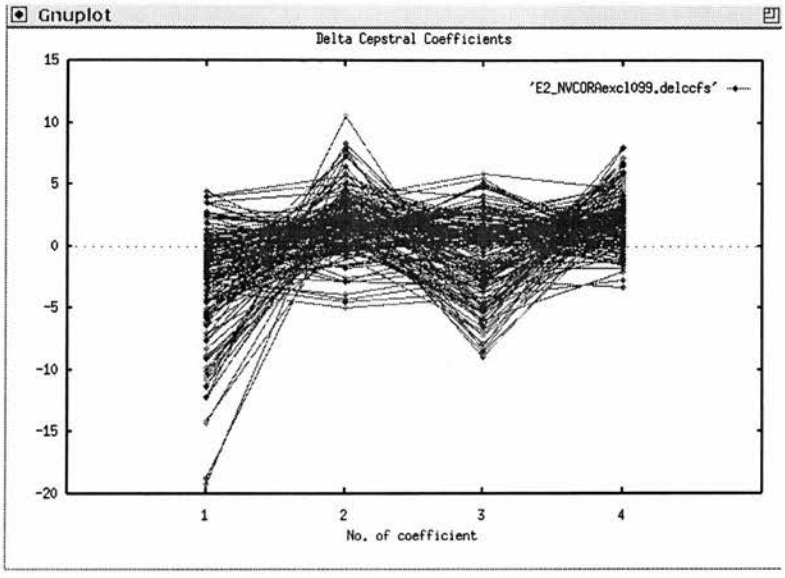



**Figure 6.27.** Training Data (dynamic coefficients) for [E2_NVCORA]

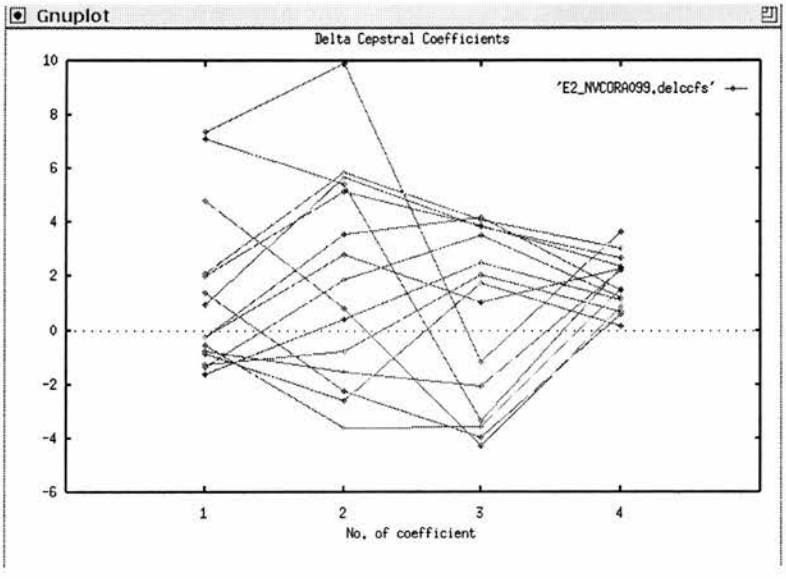



**Figure 6.28.** Test token (dynamic coefficients) [E2_NVCORA]

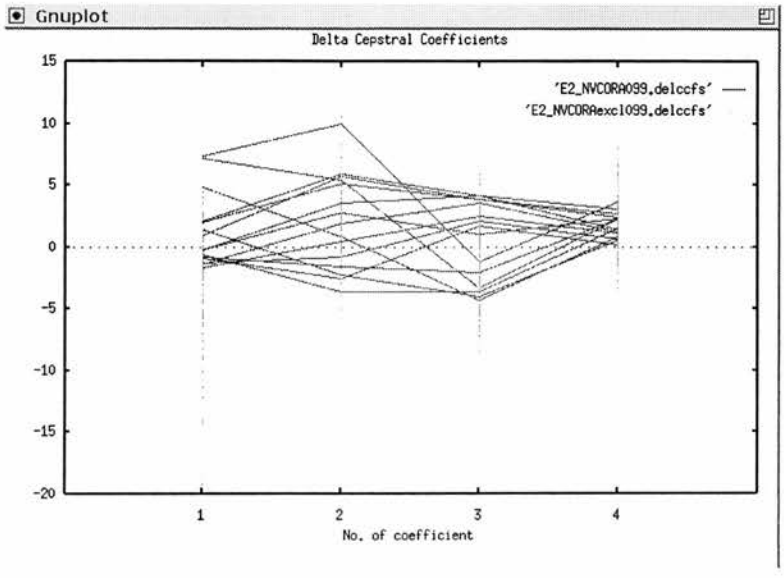



**Figure 6.29.** Combined plot of Training and Test-token data (dynamic coefficients) for [E2_NVCORA]

system with absolute duration penalties added, going from a training-set of 191 to a training-set of 252 increased the percent correct figure from 67.65 to 68.52 and the accuracy figure from 53.86 to 54.03. It should perhaps be noted also that projecting likely increases in performance with increases in training-data is complicated because of the nature of the machinery for generalisation, at least until training-data amounts ascend beyond modest levels. As an example, consider one of the problematic (dynamic) subphones, [PAL_oB] (onset of [o] following a palatal consonant). With very limited amounts of data, this class may be represented by a generalised model such as [PAL_BACKB] (onset of a back vowel following a consonant), and as data increases, the model representing it will tend to become more specific, passing through [PAL_LMB2] (onset of a low- or mid-back vowel following a palatal consonant) and eventually reaching complete specificity (within the confines of the present system) as [PAL_oB]. It is clearly not the case, then, that there is a simple linear relationship between the amount of data and the degree of 'representativeness': up to a certain point we may even be increasing the extent to which classes come to be modelled by very

| Pre-Pausal Recognition Results | | | | |
|---|---|---|---|---|
| % correct | accuracy (%) | correct | ins's | total |
| 66.12 | 42.89 | 521 | 183 | 788 |

**Table 6.26.** Results for Recognition of Final Four Phones

unrepresentative samples.

## 6.9 Proportion of Error Lying in Pre-Pausal Syllables

The particular problems of correctly recognising phones in the final syllable of an utterance, or in the final syllable before a major clause-boundary within a sentence (where the speaker may relax the tempo) are well known and have already been commented on earlier in the thesis (e.g. 2.5.9, final paragraph). Explicit marking via diacritics was there proposed as a way of dealing with the problem; one way of doing this would be to employ a marker at major clause- and sentence-boundaries of training-utterances, and to let this marker propagate a diacritic backwards to all phones up to and including the nearest vowel (or perhaps to the nearest lexically stressed vowel in polysyllabic words occurring before such a boundary), and to model all such marked phones distinctively, as liable to manifest effects associated with these locations in utterances. In this section — time having precluded the possibility of implementing this proposal — I attempt at least to quantify, if only very crudely, the proportion of the total error that is attributable to the peculiar features of phones in these locations.

Taking the HResults alignment of reference and automatic transcriptions, and considering only that fragment of each alignment that begins with the 4th phone from the end in the reference transcription, I counted the numbers of phones correct and the numbers of insertions, with the counts shown together with per cent correct and accuracy figures in table 6.26. These results may be compared with the results given earlier for recognition of entire utterances, where the per cent correct figure was 68.81 and the accuracy figure 54.94. Both correctness and

| Pre-Pausal Recognition Results, without Deltas | | | | |
|---|---|---|---|---|
| % correct | accuracy (%) | correct | ins's | total |
| 62.31 | 36.68 | 491 | 202 | 788 |

**Table 6.27.** Results for Recognition of Final Four Phones, without use of deltas

accuracy are clearly below average in these sentence-final regions, with accuracy showing the greatest decline (12.05 points) as a result of the large number of insertions.

It seemed possible that the delta coefficients, being highly sensitive to speaking-rate, could be responsible for some significant part of the poor performance at sentence-boundaries (and at major clause-boundaries) (given modelling with single gaussians). Hoping to get an approximate quantification of their share of responsibility, I ran the system without delta coefficients and repeated the examination of results for the final four phones. Results are shown in table 6.27. These figures may be compared with overall figures (running the system without deltas) of 63.54 per cent correct and 48.27% accuracy (table 6.28). The degradation evident in the terminal regions, relative to the overall figure, is 1.23 points for the per cent correct figure and 11.58 points for accuracy. These difference figures are not dramatically different from those for the same comparison using delta coefficients, but do perhaps suggest that the delta coefficients are less helpful in the terminal regions of utterances than they are in general. It certainly is the case, from inspection of the spectrographic evidence, that other factors are also significant — in particular, devoicing (either complete as typically before a [-VOICE] coda, or partial as a result of fading voice-effort or depletion of air in the lungs), perturbation of glottal period giving rise to insertions, loss of power, durational extremes, and gaps of silence between phones that are typically of longer duration than any similar gaps that might be found between pairs of phones of the given type elsewhere.

In 6.6 I acknowledged that the problem of non-normality may be particularly acute in respect of delta coefficients, and certainly the use of mixture-modelling represents another possible response to the peculiarities of the deltas in these

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 63.54 | 48.27 | 6375 | 534 | 3124 | 1532 | 10033 |

**Table 6.28.** Complete Transcription as Reference, with no use of deltas

particular utterance-locations. However, there are other reasons for separate labelling of phones in these locations, principally reasons associated with realistic duration-modelling, and there seems no reason for making mixture-modelling and distinctive labelling an either-or choice.

The full figures for results using entire reference transcriptions but without use of delta coefficients are given in table 6.28, for comparison with results elsewhere.

## 6.10 Cepstral Representations

In all the experiments described so far in this chapter, a single cepstral representation was used (this representation will be referred to as R26, and specified shortly). In this section I describe a number of alternatives to R26, and the results obtained using them. The variations tried in these alternatives were all in respect of the manner of frequency-warping and clustering of FFT powers; all the representations kept fast to the basic 16-element scheme described in Chapter 3 (10 static cepstral coefficients, dynamic coefficients for the first 4 of these, log power and its dynamic counterpart).

The first 5 representations (including R26) involve the simulation of rectangular, non-overlapping filters. The specifications and results for each of these are as given in tables 6.29 to 6.38.

1. R26

   The specifications for this representation are given in table 6.29, and results for use of this scheme with the best configuration found above are repeated for ease of reference in table 6.30.

| Specifications | | |
|---|---|---|
| frequency-range (Hz) | No. of bands | Bandwidth (Hz) |
| 30 - 2500 | 20 | 125 |
| 2530 - 3030 | 2 | 250 |
| 3060 - 4060 | 2 | 500 |
| 4090 - 6090 | 2 | 1000 |

**Table 6.29.** R26 Banding Scheme

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.81 | 54.94 | 6904 | 472 | 2657 | 1392 | 10033 |

**Table 6.30.** R26 Results

2. Bark Transformation (Bengio's approximation)
   The specifications (Bark scale) were detailed in chapter 3. The results of changing to this representation from R26 while leaving everything else unchanged are shown in table 6.31, and confidence and significance figures are given in table 6.32. The result for substitutions is highly significant and indicates that in respect of this feature at least the pure Bark scale is much less effective than the frequency-warping used in Representation 1, confirming earlier findings in more limited tests (described in chapter 3). The small difference of 8 in the number of deletions is not significant, but there are grounds for thinking that the difference of 81 in the number of insertions represents something real, even though the P-value falls within the 0.017 threshold.

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 66.80 | 52.12 | 6702 | 480 | 2851 | 1473 | 10033 |

**Table 6.31.** Bark Representation Results

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.355079 | 0.710158 |
| Substitutions | 0.000012 | 0.000024 |
| Insertions | 0.017922 | 0.035844 |

**Table 6.32.** Confidence and Significance Measures

| Specifications | | |
|---|---|---|
| frequency-range (Hz) | No. of bands | Bandwidth (Hz) |
| 220 – 2470 | 18 | 125 |
| 2500 – 3000 | 2 | 250 |
| 3030 – 4060 | 2 | 500 |
| 4090 – 6090 | 2 | 1000 |

**Table 6.33.** R24 Banding Scheme

3. R24

   The specifications for this scheme are shown in table 6.33. The results of switching to this representation while leaving everything else unchanged are shown in table 6.34. This result suggests again that a finer resolution is required than that given by the Bark transformation in the region below 2500 Hz. The principal difference from Representation R26 (which appears superior to this) is that the lowest 7 FFT powers are ignored, with the basis functions of the cosine transform being at their original value of unity at approximately 220 Hz. This does not appear from the result to be a sensible idea.

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 67.97 | 53.72 | 6819 | 474 | 2740 | 1429 | 10033 |

**Table 6.34.** R24 Results

| Specifications | | |
|---|---|---|
| frequency-range (Hz) | No. of bands | Bandwidth (Hz) |
| 30 - 1030 | 11 | 93.75 |
| 1060 - 3060 | 16 | 125 |
| 3090 - 4090 | 2 | 500 |
| 4120 - 6120 | 2 | 1000 |

**Table 6.35.** R31 Banding Scheme

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 66.79 | 52.13 | 6701 | 464 | 2868 | 1471 | 10033 |

**Table 6.36.** R31 Results

4. R31

   Specifications for this scheme are given in table 6.35 and results in table 6.36. The motivation for trying this representation was to see if better discrimination could be achieved between confusible back vowels and [dl] by (a) getting a finer resolution of the low-frequency region below 1000 Hz and (b) getting a finer resolution of the region between 2500 and 3000 Hz in order to pick up on the distinguishing feature of some [dl] tokens, that of very high (and sometimes very weak) F3. The desired results did not follow, and the generally negative effect on performance is clear from the table.

5. R23

   Specifications for this scheme are given in table 6.37 and results in table 6.38. This representation reflects a retreat from the very narrow bandwidth for the region below 1000 Hz which was tried in Representation 4, with a compromise bandwidth for the region between 1000 and 3300 Hz. Ideally, one would have postponed the switch to a 2000 Hz bandwidth for the highest frequency-region until a further experiment. The result is better than any

| Specifications | | |
|---|---|---|
| frequency-range (Hz) | No. of bands | Bandwidth (Hz) |
| 30 - 1030 | 8 | 125 |
| 1060 - 3280 | 12 | 187 |
| 3310 - 4310 | 2 | 500 |
| 4340 - 6340 | 1 | 2000 |

**Table 6.37.** R23 Banding Scheme

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 68.22 | 53.96 | 6845 | 472 | 2716 | 1431 | 10033 |

**Table 6.38.** R23 Results

of the others except the best, but it is not clear what moral, if any, can be drawn other than what has already been offered.

Having found R26 to be the best of those tried (though this hardly constitutes a well-worked-out series of experiments), I attempted to combine the same frequency-warping with an implementation of the Davis and Mermelstein scheme with overlapping triangular filters. This is described as Representation TRI33.

In Representation 1 the filters were non-overlapping, and each FFT power within a band was given equal weight. In the present scheme, the left edge of every filter other than the first falls at the centre of the filter that precedes it, and the powers are given increasing weight (in accordance with the triangle function) as they approach the power at the centre-frequency (which is given a weight of 1). In general the filters spanned larger numbers of powers than did the corresponding rectangular filters of Representation 1, but the first and last power in each case were nullified by the triangle function, and the next outermost left and right powers given relatively little weight, and for the first 26 of the 33 filters only 5 FFT powers were given non-zero weight in each case. The bandwidths in the table (6.39) need to be seen in the light of the triangular function.

The results using this representation are shown in table 6.40. Figures for

| Specifications | | | |
|---|---|---|---|
| frequency-range (Hz) | No. of bands | Bandwidth (Hz) | Centre-frequency |
| 30 - 2690 | 27 | 160 | 62, 155, 248, ... |
| 2530 - 2840 | 1 | 310 | 2690 |
| 2690 - 3190 | 1 | 500 | 2840 |
| 2840 - 3660 | 1 | 820 | 3190 |
| 3190 - 4280 | 1 | 1090 | 3660 |
| 3660 - 4900 | 1 | 1240 | 4280 |
| 4280 - 5530 | 1 | 1250 | 4900 |

**Table 6.39.** TRI33 Banding Scheme

| Phone Recognition Results | | | | | | |
|---|---|---|---|---|---|---|
| % correct | accuracy (%) | correct | del's | subst's | ins's | total |
| 65.49 | 51.84 | 6571 | 728 | 2734 | 1370 | 10033 |

**Table 6.40.** TRI33 Results

significance and confidence (with R26 as the control) are given in table 6.41. This result appears to confirm the results of much earlier, preliminary experiments, suggesting that the idea of simulating triangular filters with a view to smoothing the output representation is in fact misconceived. Certainly the worsening in the number of deletions is highly significant.

The fact that variation of the banding-scheme could have such significant effects on performance suggests that some further improvement could be found

| Confidence/Significance | | |
|---|---|---|
| Feature | c-value | P-value ($P_0 = 0.017$) |
| Deletions | 0.000002 | 0.000004 |
| Substitutions | 0.032418 | 0.064836 |
| Insertions | 0.737148 | 0.525704 |

**Table 6.41.** Confidence and Significance Measures

by further (and more systematic) experiment. It is quite possible, however, that any scheme found to be optimal might prove to be so only for the given speaker.

Banding apart, there would (had time allowed) have been scope for further experimentation to assess the effects of other departures from the basic format used in all the representations detailed above: adding further dynamic coefficients one at at a time, for example, treating static and dynamic features as independent, adding delta delta coefficients, trying additional features such as frame-to-frame correlations, and so on. (Many of these things were in fact tried informally in preliminary experiments, and not followed through because at the time the hunt was for 'dramatic' improvements, but in any case all of them have been fully investigated and reported in the research literature, and there is no particular reason for thinking that applying them in the present system would give rise to unusual results.)

## 6.11 Concluding Remarks

Many things have had to be excluded from the critical evaluation presented in this chapter, simply because of lack of time. The problem of 'gliditis' is, for example, only one of a small number of categories of pervasive error, the two most serious apart from it being (1) confusions between elements of the [ou] diphthong ([axD1 axrD2] or [axD1 uumD2]) and [lax] or 'canonical' tokens of schwa, and between elements of the [au] diphthong ([aD1 axD2] or [aD1 uumD2]) and [a], both confusions appearing to be particularly likely before nasals, and (2) confusions between any of a number of core-state back vowels on the one hand and [Cdl] on the other, or between any of these vowels and [ax dl] sequences when the vowel follows a velar, palatal or alveolar consonant. Several months of additional work might have solved these problems, though the problems surrounding [dl] are difficult.

An indication has been given of the current performance of the system (though the headline figures should ideally be looked beyond, and transcription-output from the best configuration is included as an appendix). Claims about the potential for improvement have been tentative, but I hope will be deemed reasonable.

# Chapter 7

# Conclusion

It is clearly not always possible to tell in advance, either in theoretical science or in the development of technology, whether further progress will come from continued patient work along lines already established, or from some new or radical departure from them, and it is not impossible that the ultimate solution to automatic speech recognition (where "ultimate" implies at least a robust ability to make sense of wholly spontaneous speech) will come as the result of unexpected and novel developments. However that may be, it is clearly advantageous in the meantime to work with techniques that allow as straightforward as possible identification of cause-and-effect relationships, and that allow modification and development to accommodate new knowledge as it becomes available.

If the case has been made convincingly in this thesis that subphonic analysis is possible outside the context of HMM,[1] then it may be said that the dual process ontology of HMM represents an unnecessary complication; it certainly makes conceptualisation of the recognition process difficult, and so presents an obstacle to the formation of hypotheses regarding the causes of problems and regarding their possible solutions. Explicit modelling of subphonic elements makes possible a considerable degree of simplification, and appears to open the way to a framework for transcription that can be adapted readily to take account of advances in knowledge.

[1] even if the case has been demonstrated only for the case of a single speaker

The suggestion of alternative methods for the resolution of phones into sub-phones perhaps represents the most important contribution of the thesis, in spite of its simplicity (and in spite of the fact that the implementation of the methods is probably far from optimal). On a rather different front, some of the techniques for systematic handling of difficult phonetic material, for example, for handling glottalisation, may well (it is hoped) prove to be of value as ASR begins to do more serious battle with unscripted and relatively unstructured speech, where high-level constraints are less available, or require very sophisticated formulation, so that phonetic cues to structure and to breakdowns of structure (via hesitation, false starts, long pauses, stutter, and the like) need to be taken seriously. It is also my hope that the phonetic transparency of the technique described may help to draw more people with interests centered in Phonetics and even Phonology to take a greater interest in the field of ASR. There has perhaps been a tendency for people with interests in these areas to regard the phonetic frameworks typically used in statistical ASR as crude and uninteresting, and I hope to have gone some way in Chapter 2 toward showing that an overtly "linear segmental" framework need not necessarily be as crude or uninteresting a thing as some have appeared to think.

As far as the question raised in Chapter 1 is concerned, regarding the possibility, at least for a given speaker, of achieving automatic phonetic transcription without constraining the recognition progress using lexical or syntactic constraints, only very tentative conclusions are justified, given that it is difficult to to say just *how much* improvement could be achieved in performance. (Speech recognition techniques using Hidden Markov Modelling have been undergoing intensive development for some twenty years, and it seems reasonable to claim that FURIDA would benefit from further development in the same way.) Certainly numerous ideas have been put forward earlier for attempting to secure improvements, including

- greatly increasing the amount of training-data, this also allowing use of a full complement of static and dynamic features rather than the reduced set of dynamic features used in this work;

- modelling with mixtures;

- spectral resolution of all TR classes;
- making the TR classes themselves as specific as the amount of training-data will allow;
- effecting a comprehensive separation between classes and the models used to represent them, in order to exploit the full potential of phonetic sequencing-constraints;
- explicit modelling of phones in regions preceding major clause-boundaries and sentence-boundaries;
- abandoning a number of less desirable merges, such as merges of nasal and non-nasal contexts of onsets and offsets of vowels, and merges of the initial silence context and the glottal stop context for onsets of vowels;
- attempting relativistic (within-phone) duration-modelling for subphones;
- moving away (at least in certain instances) from merely piecewise context-dependency to triphonic or similar sorts of dependency
- attempting to develop something more sophisticated than the present crude methods for subphonic resolution of most consonants;
- integrating subphone-based DP search for the best transcription within an expanded network of higher-level phonetic structures (perhaps syllables), within which to be able to score competing hypotheses taking account not only of wider phonotactic constraints, but also of more realistic durational probabilities, than are available outside such a structure.

Once such improvements had been attempted, it would begin to be clearer just how far the Phonetic Self-Sufficiency Hypothesis of Chapter 1 could continue to be maintained. Certainly there would appear even from the results presented in Chapter 6 (to say nothing of work elsewhere in ASR) to be good grounds for thinking that the weak form of the hypothesis is defensible.

It would clearly be no small or simple matter to go from the speaker-dependent to the speaker-independent case, using the technique described in this work. For

a truly speaker-independent system for a 'world language' like English, one is faced with a huge range of variation in the phonetic implementation of phonological categories and indeed in the phonological systems used to 'implement' the lexicon. To be able to reach a point where one could retain in such a context the freedom from artificial constraints that is promised by accurate phonetic classification, a great deal of thought and work would obviously be required. A variety of approaches can be envisaged, but this is a topic which takes us beyond the confines of the present work. It is in any case perhaps worth noting that speaker-independent systems, able to recognise the speech of a previously unencountered individual without warning, have their most obvious applications in a number of areas where the nature of possible exchanges is highly predictable, making higher-level constraints less objectionable, and that there is also a great deal of scope for systems of other kinds. It is easy to imagine uses for systems in which it was possible to learn a speaker's phonological system and its phonetic implementation rapidly from a brief initialisation, and then on subsequent encounters switch to the appropriate recogniser for that speaker in accordance with some form of speaker ID. This is easier to imagine if the 'socialisation' of speech recognition technology becomes an integral part of the general expansion in telecommunications and computer services. With global pooling and networking of speech-recognition resources, it is possible to imagine universal access to a host of 'baseline' accent-specific recognisers, trained in the case of widely used accents from vast amounts of data, with speaker-initialisation routines to identify the best baseline system for a given user, and further routines for tuning to that speaker's specific requirements and for longer-term adaptation to it. Adult speakers of English nowadays can typically cope with English from any part of the world (though a few accents still cause considerable difficulty to those unfamiliar with them!), and the reason is partly to do with exposure to them via TV and radio etc., and partly due to the similarities and systematic relationships between all specific forms of English and English as a more formal or abstract system. In both cases we are brought back to the starting-point of the thesis – the crucial importance of experience in making robust recognition possible for humans. There is every reason to think that providing the analogue of this for machine recognition is every bit as crucial if machine-recognition is ever to reach

the level of performance of humans.

# Appendix A

# Phone Labels

| LABEL | EXPLANATION |
|---|---|
| GL | Glottal stop |
| GL* | any phone following a non-fast glottalisation |
| GS* | Glottal stop with annotation for following vowel |
| Kb | unaspirated release of [k] as in 'sky' |
| Krb | coarticulated [Kb] and [r] as in 'scream' |
| PNASdh | [dh] following nasal |
| PNVdh | [dh] following [-VOICE] phone |
| PVdh | [dh] following [+VOICE] phone |
| Pb | unaspirated release of [p] as in 'spy' |
| Tb | unaspirated release of [t] as in 'sty' |
| Trb | coarticulated [Tb] and [r] as in 'strode' |
| a | vowel of 'bad' |
| aD1 | first element of diphthong of 'now' |
| aa | vowel of 'hard' |
| aa2ax | form of diphthongal glide of diphthong of 'my' |
| aa2e | form of diphthongal glide of diphthong of 'my' |
| aa2i | form of diphthongal glide of diphthong of 'my' |
| aa2ii | form of diphthongal glide of diphthong of 'my' |
| aaD1 | first element of diphthong of 'my' |
| ax | schwa |
| axD1 | first element of diphthong of 'go' |
| axD2 | form of second element of diphthongs of 'now' and 'here' |
| axD3 | form of third element of diphthongs of 'my' and 'boy' |
| axrD2 | form of second element of diphthongs of 'now' and 'go' |
| bb | release of [b] |
| bbc | complex closure of two successive [b]'s |
| bc | closure of [b] |
| bdc | complex closure of [b] followed by [d] |
| blb | coarticulated [bb] and lateral |
| bnc | [b] produced without closure |
| chb | release phase of [ch] as in "chew" |
| chc | closure phase of [ch] as in "chew" |
| cl | non-dark lateral with weak F2 |
| db | release of [d] |
| dbc | complex closure of [d] followed by [b] |
| dc | closure of [d] |
| dchc | complex closure of [d] followed by [ch] |
| ddc | complex closure of two successive [d]'s |
| dgc | complex closure of [d] followed by [g] |
| dh | consonant of "though" |
| dhR | release phase of [dh] given stop-like realisation |
| dhS | closure phase of [dh] given stop-like realisation |
| dl | dark lateral |
| dlo | devoiced phase of dark lateral |

| LABEL | EXPLANATION |
|---|---|
| dnc | [d] produced without closure |
| dpc | complex closure of [d] followed by [p] |
| drb | coarticulated [db] and [r] as in 'dry' |
| drc | closure of [d] occurring before [r] |
| dtc | complex closure of [d] followed by [t] |
| e | vowel of 'bed' |
| e2i | form of glide in diphthong of 'say' |
| e2ii | form of glide in diphthong of 'say' |
| eD1 | first element of diphthong of 'say' |
| eD3 | form of third element of diphthongs of 'say', 'try', 'boy' |
| eHD1 | first element of diphthong of 'care' |
| f | consonant of 'foe' |
| gb | release of [g] |
| gc | closure of [g] |
| ggc | complex closure of two successive [g]'s |
| gnc | [g] produced without closure |
| grb | coarticulated [gb] and [r] as in 'grow' |
| gtc | complex closure of [g] followed by [t] |
| h | consonant of 'he' |
| i | vowel of 'hit' |
| iD3 | form of third element of diphthongs of 'say', 'try', 'boy' |
| ii | vowel of 'see' |
| iiD1 | first element of diphthong of 'near' |
| iiD3 | form of third element of diphthongs of 'say', 'try', 'boy' |
| ir | reduced form of [i] or fronted schwa |
| irD2 | form of second element of diphthong of 'fear' |
| jhb | release of [jh] as in 'join' |
| jhc | closure of [jh] as in 'join' |
| kb | release of [k] |
| kc | closure of [k] |
| klb | coarticulated [kb] and lateral as in 'Clara' |
| krb | coarticulated [kb] and [r] as in 'cry' |
| ktc | complex closure of [k] followed by [t] |
| l | non-dark lateral |
| lax | vowel of 'bird' |
| lo | voiceless non-dark lateral |
| m | consonant of 'my' |
| mo | devoiced [m] |
| msyl | syllabic [m] |
| n | consonant of 'now' |
| ng | consonant of 'wrong' |
| no | devoiced [n] |
| nsyl | syllabic [n] |

| LABEL | EXPLANATION |
|---|---|
| o | vowel of 'hot' |
| oe | open form of [e] occurring before [dl] or after [w] |
| oo | vowel of 'caught' |
| oo2ax | form of glide of diphthong of 'boy' |
| oo2i | form of glide of diphthong of 'boy' |
| oo2ii | form of glide of diphthong of 'boy' |
| ooD1 | first element of diphthong of 'boy' |
| pb | release of [p] |
| pbc | complex closure of [p] followed by [b] |
| pc | closure of [p] |
| plb | coarticulated release of [p] and lateral as in 'play' |
| prb | coarticulated release of [p] and [r] as in 'pray' |
| ptc | complex closure of [p] followed by [t] |
| r1 | first phase of [r] (falling F3) |
| r2 | second phase of [r] (rising F3) |
| rii | vowel of 'ream' occurring after [r] |
| riiD1 | first element of diphthong of 'rear' occurring after [r] |
| s | consonant of 'so' |
| sh | consonant of 'show' |
| tb | release of [t] |
| tc | closure of [t] |
| tflap | flapped [t] |
| th | consonant of 'thigh' |
| thR | release of [th] given a stop-like realisation |
| thS | closure of [th] given a stop-like realisation |
| thr | coarticulated [th] and [r] as in 'thread' |
| tkc | complex closure of [t] followed by [k] |
| tlb | coarticulated release of [t] and lateral as in 'at last' |
| tpc | complex closure of [t] followed by [p] |
| trb | coarticulated release of [t] and [r] as in 'try' |
| trc | closure of [t] occurring before [r] |
| ttc | complex closure of two successive [t]'s |
| u | vowel of 'put' |
| uD1 | first element of diphthong of 'poor' |
| uh | vowel of 'trust' |
| ukc | closure of unaspirated [k] |
| upc | closure of unaspirated [p] |
| utc | closure of unaspirated [t] |

| LABEL | EXPLANATION |
|---|---|
| uu | form of vowel of 'school' |
| uuf | form of vowel of 'you' |
| uum | form of vowel of 'move' |
| uumD2 | form of second element of diphthongs of 'now', 'go' |
| v | consonant of 'vie' |
| w1 | first element of [w] (weakening F3, falling F2) |
| w2 | second element of [w] (strengthening F3, rising F2) |
| y | consonant of 'yeah' |
| z | consonant of 'zee' |
| zh | first sound of 'genre' |

# Appendix B

# HResults Aligned Transcriptions

Transcriptions are given just for the first 100 sentences.

```
sc001.auto:  63.41(41.46)  [H=26, D=2, S=13, I=9, N=41]
Aligned transcription: sc001.hand vs sc001.auto
 LAB: dh ax    pc prb r      AIgl D3 s r     EIgl D3 n   zh  i  z s m
     oo   l ax  dh  ax n e n ii v ax s    ax kc   kb s upc Pb e  GLsc
          tb ir dc db
 REC: dh ax bc pc prb r aaD1 AIgl D3 s r eD1 EIgl ng jhc jhb ir z s m
ooD1 OIgl l lax dnc ax n a n i  w ax s tb y  GLsc kb s        ax GLsc
bb a kc tc tb ir dc db
sc002.auto:  74.42(62.79)  [H=32, D=0, S=11, I=5, N=43]
Aligned transcription: sc002.hand vs sc002.auto
 LAB: dh EIgl D3 aa s ukc Kb Tb ir f      AIgl    w o  n tc tb ir dc tc
tb ir kc kb uh m   ax l  o ng o n dh ax bc bb          aa jhc jhb trb
r ir GLsc pb
 REC: bb EIgl D3 aa s ukc kb dh ir f aaD1 AIgl D3 w uh n tc tb ir dc tc
tb ir kc kb uh m w ax dl o m  o n dh ax bc bb aaD1 AIgl D3 dnc chb trb
r ax GLsc kb
sc003.auto:  69.44(61.11)  [H=25, D=3, S=8, I=3, N=36]
Aligned transcription: sc003.hand vs sc003.auto
 LAB: ax m uh   ng s utc Tb ax f r     e    n zh sh  uuf     ir z kc
kb ax n s ir dnc ir dc bc bb y uuf tc tb ax f u  dl
```

```
 REC: uh m uh m kc s utc Tb ax f r eD1 EIgl n    chb uuf uum ax z ukc
Kb ax n s        ir dc bc bb y uuf tc tb ax f oo dl
sc004.auto:  68.33(51.67)  [H=41, D=3, S=16, I=10, N=60]
Aligned transcription: sc004.hand vs sc004.auto
 LAB: dh ax s m oe dl ax v  dh  ax f r e sh lo l ii gc grb r aD1 axrD2
n kc kb o    f ii n e v ax  f EIgl D3     dl z    tc tb uum  i  n tc tb
aaD1 AIgl D3 s    m ii i n    tb ax   dh ir sh       o pc   pb
 REC: dh ax s n ax l  ax dl dnc ax f r e sh lo   i  kc krb r     ax
n kc kb o th f ii n e v lax f AIgl D3 o w ax z db tc tb axD1 ir n tc tb
aaD1 AIgl D3 s pc m ii   n tc tb ax v dh ax sh ax dl o GLsc pb tc
sc005.auto:  71.19(57.63)  [H=42, D=3, S=14, I=8, N=59]
Aligned transcription: sc005.hand vs sc005.auto
 LAB: aa m o f    msyl pc pb ax    pc pb  l e GLsc kb s bc bb aaD1 AIgl
   r a pc   pb  ir dc db ir    dc db v aa n s ir z ir n   s Tb EIgl D3
tc tb ir v dh   ii aa    GLsc tb e    GLsc n  o l ax jhc jhb    ii
 REC: uh m o f ax m    bc bb ax pc pc plb l e GLsc kb s bc bb aaD1 AIgl
D3 r a GLsc dhR ir dc db ir dc dc db   aa n s ax z ir n z s Tb EIgl
dc db ir   GLdh ii aa pc tc   tb e kc tc   tb o l ax jhc jhb ii ii
sc006.auto:  64.10(51.28)  [H=25, D=1, S=13, I=5, N=39]
Aligned transcription: sc006.hand vs sc006.auto
 LAB: jhc jhb     o n kc kb u dc db l e  n db i m dh ax l EIgl D3 tc tb
ir z drc drb r aa f utc Tb  ax v ir z w      lax GLsc kb
 REC: jhc jhb uum o n kc kb u dc db l ax n db i n db ax l      e  dc db
i  z tc  trb r aa f tc  dhR ax v ir z w oo l ax  kc   kb tc tb
sc007.auto:  61.70(53.19)  [H=29, D=2, S=16, I=4, N=47]
Aligned transcription: sc007.hand vs sc007.auto
 LAB: f r    m    f oo tb ii l   uh  v  dh  ax s ukc Kb oo w ax z
n    aD1 axD2 jhc jhb   uuf s   ax n dh ax kc krb r aD1 axrD2 dc db gc
grb r    uuf tc tb e n NO   s
 REC: f r ax m ax f oo tb i  GLl lax dl dnc ax s ukc Kb oo      z db
nsyl aD1 axD2 n   GLchb uuf GLs ax n dh ax kc krb r aD1 axD2  pc pb kc
krb r ax kc  tc tb e n GLsc s
sc008.auto:  60.00(45.71)  [H=42, D=3, S=25, I=10, N=70]
```

```
Aligned transcription: sc008.hand vs sc008.auto
 LAB: dh ax    pc prb r e  z bnc ax tc tb      ii   r rii ir n m i n
ir s utc Tb ax m     a    n i zh tc tb ir kc kb lax       bc dh ax drc
drb r i ng kc kb i  ng   h a bc bb ir tc s   ax v dh ax l    ooD1 OIgl
trc trb r i   ng y  uuf th       s
 REC: dh ax pc pc prb r ax z l   ax tc tb iiD1 axD2 r rii    i m e n ax
z  s utc Tb ax m aD1 axD2 l i s  tc tb ir kc kb lax dc dc db ir dc tc
trb r e n  kc kb ii ng i h a bc bb    i  GLs ax v    ax l ax w    ax
tc  trb r rii ng ii ir  th pc pb s
sc009.auto:  61.29(45.16)  [H=19, D=1, S=11, I=5, N=31]
Aligned transcription: sc009.hand vs sc009.auto
 LAB: dh ax    bc bb uh dl bc bc bb l     uum    w oe n ii s w ir GLchb
utc Tb o n dh ax l    aaD1 AIgl D3 tc tb
 REC: bb ax bc bc bb    o  bc bc bb l uuf uum ax w ax l ii s w ax GLsh
dc  db o n dh ax l aa h    i    tc tc tb ax
sc010.auto:  75.00(58.33)  [H=27, D=1, S=8, I=6, N=36]
Aligned transcription: sc010.hand vs sc010.auto
 LAB: i  tc tb ir z         f y uuf tc tb aa dl tc tb    u o f  r e  n
ii       f lax dh ax r ir z i  s utc Tb ax n NO   s
 REC: ir tc tb ir z s tc tb h y uuf tc tb aa    tc tb ax w o tb r ir h
ii dc th f lax dh ax r ir z ir s utc Tb ax n GLsc s
sc011.auto:  73.17(63.41)  [H=30, D=1, S=10, I=4, N=41]
Aligned transcription: sc011.hand vs sc011.auto
 LAB: dh   EIgl l oo n GLchb Tb i  n tc tb ax    bc bb a       tlb dl w
ir dh oo dl dh ax f oo s ir z dh EIgl kc kb ir dc m uh s utc Tb ax
 REC: dh i dnc  l oo n chb   Tb ax n tc tb ax bc bc bb a th lo l   ax w
ax z  oo    tb ax f oo s ir z db EIgl kc kb ir dc m aa s utc Tb ax
sc012.auto:  67.65(55.88)  [H=23, D=1, S=10, I=4, N=34]
Aligned transcription: sc012.hand vs sc012.auto
 LAB: dh ir chc chb    i dl w i n dc db kc kb oo z dh ax m tc tb ir sh
i  v ax v aaD1 AIgl    l ax n   tlb l      ii
 REC: dh ax chc chb ax l ax w i n    tb kc kb oo z dh ax n tc tb ax sh
ir v ax v dl   AIgl D3 l ax n s lo  l EIgl D3
```

```
sc013.auto:  70.27(58.11)  [H=52, D=1, S=21, I=9, N=74]
Aligned transcription: sc013.hand vs sc013.auto
 LAB:    dh ir gc gb     uh v ax m ax n GLsc trb r aaD1 AIgl D3 ax m
GLsc tb f      oo   y  i  ir z ir gc gb axD1 axrD2 nsyl w    ii ax  v
     e    v  r rii    r rii        z  nsyl tc tb ax bc bb ax l ii v dh
ax tflap i  GLw w  dl trc trb r aaD1 AIgl    ax m NO f ax gc gb
e    n
 REC: tc dh i  gc gb uuf o  v ax m ax n trc  trb r aaD1 AIgl D3 ax
GLsc tb f ooD1 OIgl D3 ii ir z ir gc gb axD1 axrD2 m    AIgl D3 lax v
eHD1 axD2 th r rii ax r rii zh dnc ax n    tc tb ax bc bb ax l ii v dh
ax h     ir GLw oo dl tc  trb r aaD1 AIgl D3 ax n tc f ax gc gb eHD1
axD2 n
sc014.auto:  71.43(61.90)  [H=30, D=3, S=9, I=4, N=42]
Aligned transcription: sc014.hand vs sc014.auto
 LAB: h ii jhc  jhb      lax   kc tc Trb r aD1 axrD2 n db ir n ax n i n
s utc Tb ax n tc tc tb ax f      EIgl D3 s ir z ax s EIgl D3 l ax
NO   tb
 REC: h ii gjhc jhb axD1 axrD2 kc tc trb r aD1 axD2  n    ir      n i n
s tc  tb ax n tc tc tb ax f lo l EIgl D3 s ir z ax s lo   i  l ax n
GLsc tb
sc015.auto:  72.73(70.45)  [H=32, D=0, S=12, I=1, N=44]
Aligned transcription: sc015.hand vs sc015.auto
 LAB: h ii e  m  f ax s aaD1 AIgl D3 z ir z s Trb r      e  ng GLsc kb
s w aaD1 AIgl kc kb ax n s ii l i ng i z w ii kc kb n  ax s ir z
 REC: h ii ax tc f ax s aaD1 AIgl D3 z ir z s Trb r EIgl D3 n  pc   pb
s w uh   pc   kc kb ax n s ii l i ng i z w i  kc tc Tb ax s ir s
sc016.auto:  70.27(62.16)  [H=26, D=4, S=7, I=3, N=37]
Aligned transcription: sc016.hand vs sc016.auto
 LAB: dh ax tc tb     EIgl D3 bnc l ir z m       EIgl D3 dc db s axD1
uumD2 s lo l o pc pb  l ii dh ir tflap ir GLsc tb ir dl tc tb   s
 REC: dh ax tc tb eD1 EIgl D3 bnc l ir z m i GLn EIgl D3 i  z  s axD1
uumD2   z  l o pc plb l ii       dnc   ir GLsc tb e  dl    GLsc s
sc017.auto:  72.09(60.47)  [H=31, D=5, S=7, I=5, N=43]
```

```
Aligned transcription: sc017.hand vs sc017.auto
 LAB: i GLw w ir z    ir m pc pb oo    tc nsyl tc tb ax bc bb  ii    pc
pb lax f ir GLsc s ax n s      dh  ax w ax n axD1 uumD2    pc prb r o m
GLsc tb s
 REC: i     w ax z db ir n pc pb oo dc db nsyl    tb ax    GLm ii kc pc
pb lax f i  GLsc s ax n s TDHS dhR ax w ax n axD1 uumD2 pc pc prb r o
     NO s
sc018.auto:  67.65(52.94)  [H=23, D=1, S=10, I=5, N=34]
Aligned transcription: sc018.hand vs sc018.auto
 LAB:      aa   r a n    f ax kc kb uh v ax w aa s utc Tb ii h lax dl
dc db   s ax v r dl s utc Tb  axD1 axrD2         n z
 REC: aaD1 AIgl r a n tc f ax kc kb uh v ax w uh s dc  db ii h lax l
oo GLsc s ax v   dl s tc  dhR axD1 axrD2 ax n ax n s
sc019.auto:  63.16(44.74)  [H=48, D=1, S=27, I=14, N=76]
Aligned transcription: sc019.hand vs sc019.auto
 LAB:    w  ii ax v  pc prb r uuf f    dh  ax    GLsc dh ax r ir  zh ii
   m w ii dl dc   s ax f i   sh nsyl                  GLsc pb a  r ir
n dh ax n oo    th tc tb u    ir kc kb s pc    pb l ooD1 OIgl D3 GLsc
   dh ii i n tc tb aaD1 AIgl D3 ax pc pb o pc    pb ax l EIgl D3 sh ax
n
 REC: tc gc gb e  pc pc prb r i   f tc dhR ax pc tc   Tb ax r rii zh ii
ng n w ir dl GLsc s ax f uuf sh nsyl tc dc db ax tb aD1 axD2 r  ax r i
n dh ax n oo tb th tc f  axD1 ax kc kb s pc bc bb l ax   w    i  GLsc
kb h  ii i n tc tb aaD1 AIgl D3    pc pb o pc bc bb ir l EIgl D3 sh ax
n
sc020.auto:  64.71(38.24)  [H=22, D=3, S=9, I=9, N=34]
Aligned transcription: sc020.hand vs sc020.auto
 LAB: sh ii    f  lo l i kc kb s    th r uum ax m a    gnc ir z   ii n
w oe n     sh  ii    gnc e GLs ax chc chb         aa   NO   s
 REC: sh ii tb th lo   i kc kb s tb ax r     ax m a gc gb  i  z i ng n
  ax n chc chb ii gc gb  e GLs ax chc chb ax GLAA aa n GLsc s
sc021.auto:  56.41(51.28)  [H=22, D=3, S=14, I=2, N=39]
Aligned transcription: sc021.hand vs sc021.auto
```

```
 LAB: th a    ng gc gb u  dc  n  ax s ir GLs f r aaD1 AIgl D3 dc db
EIgl    ax n tc tb aaD1 AIgl D3 m tc tb ir gc gb axD1 axrD2 h     axD1
uumD2 m
 REC: bb AIgl dc gc gb ax GLm ir z  s ir z   f r aaD1 AIgl D3 dc db
EIgl D3 e  n tc tb aaD1 AIgl      n  tb ir gc gb      lax   h lax tflap
ax    n
sc022.auto:  63.64(45.45)  [H=21, D=1, S=11, I=6, N=33]
Aligned transcription: sc022.hand vs sc022.auto
 LAB:    i  chc chb i z ax r oo dl w ir z s axD1 uumD2    tc tb e   m
utc Tb i  ng tc tb ax s kc Krb r a       GLsc    chb
 REC: pb ii chc chb i s ax r    ax w ax z s axD1 uumD2 tc tc tb e n tc
dc  db ii n  tc tb ax s kc krb r a tb ax GLsc sh tb
sc023.auto:  74.07(55.56)  [H=20, D=1, S=6, I=5, N=27]
Aligned transcription: sc023.hand vs sc023.auto
 LAB: aa dl   h  e jhb m aaD1 AIgl       bc bb e GLs ax n tc tb EIgl
kc n axD1 axrD2 r i      s kc  kb s
 REC: o  GLsc pb e zh  m aaD1 AIgl D3 bc bc bb e GLs ax n tc tb EIgl D3
kc n      ax    r i h ax s ukc kb s
sc024.auto:  84.62(69.23)  [H=33, D=0, S=6, I=6, N=39]
Aligned transcription: sc024.hand vs sc024.auto
 LAB: dh ax l      e  ng th ax v ax s ukc Kb            lax GLsc    kc
kb oo dc dh ax pc    pb aa s ir z bc bb aaD1 AIgl D3 tc   tb ir s utc
Tb eHD1 axD2
 REC: dh ax l EIgl D3 ng th ax v ax s ukc Kb axD1 axrD2 ax  GLsc db kc
kb oo dc db ax pc bc bb aa s ax z bc bb aaD1 AIgl D3 GLsc tb ir s utc
Tb eHD1 axD2 a
sc025.auto:  71.05(55.26)  [H=27, D=1, S=10, I=6, N=38]
Aligned transcription: sc025.hand vs sc025.auto
 LAB:           AIgl D3 oo w ir z s ii m tc tb ax f o  l ax m AIgl GLi
n s utc Tb i ng    s r aa dh ax dh ax n    r rii    z ax n
 REC: db bc aaD1 AIgl D3 ax w    i s ii n tc tb ir f lo l ax m AIgl D3
n s dc  db i ng gc s r aa dl a  dh ax n ax r rii ii z ax n
sc026.auto:  62.75(50.98)  [H=32, D=7, S=12, I=6, N=51]
```

```
Aligned transcription: sc026.hand vs sc026.auto
 LAB: i GLs trc Trb r     EIgl D3 n    zh dh ax tflap aaD1 AIgl D3 s lo
l e    GLsc f  ax    s axD1 axrD2 l o ng s ax n s AIgl D3 w ax z nsyl
GLsc f ii l ir ng      tc tb aaD1 AIgl D3 dc   db
 REC: i GLs tc  trb r eD1 EIgl D3 n ir z  dh ax            l    ax s lo
l e pc tc   Tb ax z s      lax   l o n  s ax n s      ax w ax z nsyl
     f      ir ng m pc tc tb aaD1 AIgl D3 GLsc tb
sc027.auto:  55.81(48.84)  [H=24, D=2, S=17, I=3, N=43]
Aligned transcription: sc027.hand vs sc027.auto
 LAB: aa dl drc drb r   aa f tc dh   axD1 uumD2 z    n y  uuf pc pb r
pc pb axD1 uumD2   z    dl z  bc bb ax f oo dh ax n e kc kb s m ii tc
tb i  ng
 REC:    o  n   trc trb aa   th TDHS dhR  ax    z db n ii bc  bc bb r
pc pb axD1 uumD2 u z oo tb th bc bb ax f oo dh ax n i kc kb s m ii tc
tb ii n
sc028.auto:  64.29(48.21)  [H=36, D=3, S=17, I=9, N=56]
Aligned transcription: sc028.hand vs sc028.auto
 LAB: dh ax m          uh dc   s     kc  kb w  oe dl GLchb tc tb l aD1
axrD2 dc db l ii   ax n ii r  riiD1 axD2 l aaD1 AIgl D3 z dh ax tflap
ir z s w     EIgl D3 dc bc bb uuf tc  tb   s w ax dc   db  uuf        m
dc db
 REC: dh ax m aaD1 AIgl D3 GLsc s ukc klb l  ax l  u  GLs   tc tb l aD1
axrD2 dc db l ii h i  n ii ax v     ir   l      AIgl D3 z dh
ir z s w eD1 EIgl D3 dc bc bb uuf uum GLsc s w ax GLsc dhR uuf uum ax n
dc db ax tb
sc029.auto:  69.09(63.64)  [H=38, D=3, S=14, I=3, N=55]
Aligned transcription: sc029.hand vs sc029.auto
 LAB: h ii gc grb r aa s pc  pb dh ax r axD1 uumD2 pc pb w ir dh  ir z
f r rii    h a n db ax n s w uh ng i  z l e    gc gb z r     a    n tc
tb ax dh ax s aaD1 AIgl   D3 dc db
 REC: h ii gc gb  r aa s upc pb dh ax r axD1 axrD2 pc      pc plb ir z
f r rii ii h a n db ax n s w ax n  ii z l EIgl gc    s r aD1 axD2 n tc
tb ax dh ax s aaD1 AIgl h i  tc tb
```

```
sc030.auto:  60.38(49.06)  [H=32, D=2, S=19, I=6, N=53]
Aligned transcription: sc030.hand vs sc030.auto
 LAB:    o pc pb ax   chc chb uuf ax n tb ii z l  aaD1 AIgl D3       dh
i s dc  db axD1 axrD2 n GLsc grb r    axD1 axrD2 o    n trc trb r rii
   z ax z   y  sh ir dc n axD1 uumD2 bc bb aaD1 AIgl D3 n    aD1 axrD2

 REC: bb o bc bb AIgl chc chb        i tb ii z db aaD1 AIgl D3 dc dc db
i s utc Tb lax  dl    n kc   krb r ax w    aaD1  AIgl n trc trb r rii
ii z ax dnc ii sh ir tc n axD1 uumD2 bc bb aaD1 AIgl D3 n db aD1 axD2

sc031.auto:  66.67(53.70)  [H=36, D=3, S=15, I=7, N=54]
Aligned transcription: sc031.hand vs sc031.auto
 LAB: i  GLs dc  db i  v   ax vknc uh dl tc tc tb ax chc chb uuf    z
bc bb ax tc tb   w  ii n tc tb uuf    s uh GLchb   GL ii kc kb w    l
ii gc gb u dnc o  dl tc tb lax n ax    tb ir   v  z
 REC: ir z   utc Tb ir dnc ax      bc bb oo tc tb ax chc chb uuf ax z
bc bb i  tc tb h ir ng n tc tb uuf zh s uh GLchb i h  ii kc kb w ax l
ii gc gb       ax dl tc tb lax n ax tc tb ir h ax z
sc032.auto:  57.45(40.43)  [H=27, D=2, S=18, I=8, N=47]
Aligned transcription: sc032.hand vs sc032.auto
 LAB: h      e n  r rii n    oo m   l ii v y   uuf       dc r o bc bb z
e  l ax s ax n       th     y uuf z ii a   z ax m w ax th dc   db  i  s
dc  db eD1 EIgl D3 n
 REC: h eHD1 z ax r rii n ax dl dnc l ii   GLy uuf gc bc bb r o bc    z
ax l ax s ir n dc db th i h y uuf z ii lax z ir n w ir th TDHS dhR ir z
utc Tb eD1 EIgl D3 n
sc033.auto:  53.33(42.22)  [H=24, D=0, S=21, I=5, N=45]
Aligned transcription: sc033.hand vs sc033.auto
 LAB:      sh ax   dl th i ng kb ax v ax n    i  kc kb s ukc Kb  y uuf
s   ir f  gc gb i v ax n ax n uh f  s upc Pb EIgl D3 s    nsyl tc tb
aaD1 AIgl D3 m
 REC: s tc sh ax v db ii n kc kb ax v ax n ii ng gc gb s ukc klb l uuf
GLs ir th kc kb i v ir n ax n aa tb s bc  bb EIgl D3 s ax n    tc tb
```

```
aaD1 AIgl ax n
sc034.auto:  64.71(50.98)  [H=33, D=1, S=17, I=7, N=51]
Aligned transcription: sc034.hand vs sc034.auto
 LAB: w oe n f oo s tc  tb ax m EIgl D3 kc kb ir chc chb ooD1 OIgl D3 s
     e    r ax chc chb axD1 uumD2   z  pc pb i ng pc pb o ng ax z    ax
f   EIgl D3 v r  GLsc    gb eD1 EIgl D3      m
 REC: w oe m f oo s utc Tb ax m EIgl    gc gb ir chc chb ax   w    ax s
eHD1 axD2 r ax chc chb axD1 uumD2 s pc pc pb i ng bc bb o n  ax z dh ax
f r EIgl D3 v ax GLsc kb ax v   EIgl D3 n ax n
sc035.auto:  63.93(49.18)  [H=39, D=1, S=21, I=9, N=61]
Aligned transcription: sc035.hand vs sc035.auto
 LAB: i  tc   s    ax    sh  EIgl D3 m dh ax    GLsc aa kc kb ir tc tb
e kc   kb s dc db ir z aaD1 AIgl    n f   ax      dh  ax m s    oe dl v
z ax n o GLsc f ax dh ir jhc jhb      e    n r         dl pc pb uh bc
blb l  ir kc   kb
 REC: ax GLsc s ir GLchb eD1 EIgl i  n dh ax bc bb   aa kc kb ax tc tb
e GLsc kb s tc tb ir z aaD1 AIgl D3 n GLf ax TDHS dhR ax n s Tb a  dl n
z ax n o GLsc f ax dh ir jhc jhb eHD1 axD2 n r ax w ax pc pc pb uh bc
    bb ax GLsc kb
sc036.auto:  69.39(63.27)  [H=34, D=5, S=10, I=3, N=49]
Aligned transcription: sc036.hand vs sc036.auto
 LAB: kb axD1      uumD2 bc bb ii tc tb i m  tc tb ax dh ax  l aaD1
AIgl D3 n bb aaD1 AIgl D3 f aaD1 AIgl D3 v th aD1 axrD2 z   ax n GLsc
th s   ax v ax s  e kc kb ax n dc db
 REC: kb axD1 axrD2 gc    bc bb ii tc tb   ir n  tb ax    dnc l aaD1
AIgl    n m  aaD1 AIgl D3 f aaD1 AIgl D3 f th aD1 axD2  z s ax n GLsc
th s f ax   z  Tb e kc kb ax n    tb
sc037.auto:  75.00(60.00)  [H=30, D=1, S=9, I=6, N=40]
Aligned transcription: sc037.hand vs sc037.auto
 LAB:    dh ax bc bb aa th pc plb l uh gc gb i  z m i  s i  ng s
    ax y ax dl h     a    v tc tb ir tc tb EIgl kc kb ir sh aD1 axrD2
         ax
 REC: pc th ax bc bb aa th pc plb l aa gc gb ir z m ir s ii ng s eHD1
```

```
axD2 ax l ax    h aD1 axD2 f tc tb ir tc tb EIgl kc kb ir sh aD1 axrD2
aaD1 AIgl D3
sc038.auto:  58.82(44.12)  [H=20, D=1, S=13, I=5, N=34]
Aligned transcription: sc038.hand vs sc038.auto
 LAB:   y  uu  oo GLsc tb ax   bc bb r uh   sh y ax    tc tb ii th bc
bb ax f oo   y ir gnc ax  tc tb ax bc bb   e dc db
 REC: h ii bnc oo GLsc tb ax m bc bb r AIgl sh   ax tc tc tb ii th tc
dh ax f oo w i tb ii  uum tc tb ir bc bb i n tc tb
sc039.auto:  77.42(67.74)  [H=24, D=2, S=5, I=3, N=31]
Aligned transcription: sc039.hand vs sc039.auto
 LAB: dh ii lax th   y uuf s  tc  tb ax bc bb ii f       l a GLsc bb ax
GLsc n aD1 axrD2 ax tc s ax   s f iiD1 axD2
 REC: dh ii lax th h y uuf zh utc Tb ax bc bb ii f th lo l a GLsc bb ax
tc   n aD1 axrD2 ax      GLsc s f iiD1 axD2
sc040.auto:  77.14(48.57)  [H=27, D=0, S=8, I=10, N=35]
Aligned transcription: sc040.hand vs sc040.auto
 LAB:          aa w ir sh ii   dc db aa dh ir gc grb r ax    bc bb
iiD1 axD2 dc db ax    sh eD1 EIgl D3 v  ir z m ax s utc Tb   aa sh

 REC: aaD1 AIgl D3 w ir sh ii z ax l  ax dh ax gc gb  r ax bc bc bb
iiD1 axD2 dc db ax zh sh eD1 EIgl D3 bc ir z m ax s utc Tb o tc sh ax
zh utc tb
sc041.auto:  56.82(38.64)  [H=25, D=4, S=15, I=8, N=44]
Aligned transcription: sc041.hand vs sc041.auto
 LAB: jhc jhb uuf dnc ir th f       a n dh ax m a n y  ir s ukc Kb ir
GLs w EIgl D3 tc tb i ng   f   oo h ax r      o    n dh ax    pc pb ii
    a    n          axD1 axrD2
 REC: jhc jhb uuf dnc    i  f Trb r a n dh ax m a   ng i  s kc  kb ir
GLs w         i  tb i ng n GLf ax w ax r aaD1 AIgl n dh ax bc i  h  ii
aD1 axD2 n aD1 axD2 GLsc tb
sc042.auto:  67.57(56.76)  [H=25, D=0, S=12, I=4, N=37]
Aligned transcription: sc042.hand vs sc042.auto
 LAB: dh i h  y uuf jhc jhb kc kb aa s    dlo dl w ax z    nsyl s lax
```

```
kc kb dl dc bc bb aaD1 AIgl    ax dc db ii pc m      axD1 axrD2 tc   tb

 REC: dh i tb y uuf dnc chb kc kb aa s lo l   ax w ax z ax n    s lax
kc kb oo dc bc bb aaD1 AIgl D3 dc dc db ii kc m aaD1 AIgl ax    GLsc tb

sc043.auto:  72.92(68.75)  [H=35, D=2, S=11, I=2, N=48]
Aligned transcription: sc043.hand vs sc043.auto
 LAB: jhc jhb eD1 EIgl D3 n ir dc db oo dc m a th s ax n f r e n GLsc
chb bc bb ax GLsc h  EIgl D3 tc tb ir dc dh ax r e s utc Tb ax v s ukc
     Kb uu dl
 REC: jhc jhb     EIgl D3 n i  dc db oo dc m a th s ax n f r e n tc
sh  bc bb ax GLsc tb EIgl D3 tc tb i  dc db ax r e s tc  tb ax   s ukc
kb w ax w  ax
sc044.auto:  72.92(62.50)  [H=35, D=3, S=10, I=5, N=48]
Aligned transcription: sc044.hand vs sc044.auto
 LAB: m a sh pc pb ax tc tb EIgl D3 tb axD1 axrD2 z ax m oo f   a tc n
i ng dh ax n      aa   dh ax    bc bb ooD1 OIgl D3 dl dc db ax    bc bb
EIgl    kc tc tb w uh n z
 REC: m a sh bc bb ax tc tb EIgl D3 tb      ir    z ax m oo f r a tc n
i ng n  ax n aaD1 AIgl dh ax bc bc bb ooD1 OIgl       oo tb ax bc bc bb
EIgl D3 gc tc tb w ax n s
sc045.auto:  68.97(62.07)  [H=40, D=4, S=14, I=4, N=58]
Aligned transcription: sc045.hand vs sc045.auto
 LAB: dh ax w lax dl dnc ir z bc bb i kc kb ax m i ng i  n kc krb r rii
s i ng    l ii dc db eD1 EIgl D3 n jhc jhb ir r ax s bc bb ax GLh aa l
   ii e n ii w uh n kc       kb eHD1 axD2 z
 REC: bb ax w     ax l   ir z bc bb i kc kb ax n i ng ax m kc krb   i
s i ng ax m ii dc db eD1 EIgl D3 n         ii r ax s pc pb ax h   aa l
ii ii e n ii w uh n kc klb l ii eHD1 axD2 s
sc046.auto:  65.96(51.06)  [H=31, D=3, S=13, I=7, N=47]
Aligned transcription: sc046.hand vs sc046.auto
 LAB: pb a trc trb   i GLsc kb w  ax n ii dc s   upc Pb ii        GLchb
th e r       pc pb ii bc bb ax kc kb ir z ax v     ir z kc klb l e f tc
```

```
pc pb       a  l ax GLsc tb
 REC: pb a tc  trb r i GLsc    bb oo n ii n  GLs bc  bb ii jhb db tc
tb e r ax bc pc pb ii bc bb ax    h  ir z ax tflap ir z kc klb l e f
pc pb tc tb aa n ax GLsc tb
sc047.auto:  64.10(58.97)  [H=25, D=1, S=13, I=2, N=39]
Aligned transcription: sc047.hand vs sc047.auto
 LAB: dh ii z  pc prb r a tc tb kc  kb dl jhc jhb axD1 axrD2 kc kb s ax
bc bb ii ng tc tb EIgl    kc kb ax n m uh GLchb tc tb uuf   f aa
 REC: dh ii zh pc prb r a pc f  ukc Kb oo jhc jhb axD1 axrD2 gc gb s ax
bc bb ii    n  tb EIgl D3 gc gb ax n m aa s     tc tb uuf v f aa
sc048.auto:  77.78(52.78)  [H=28, D=1, S=7, I=9, N=36]
Aligned transcription: sc048.hand vs sc048.auto
 LAB: aaD1 AIgl sh   dl    pc pb e n  GLsc dh ir s r u  m      axD1
         uumD2 v w ax dh ir f   y uuf    bc bb     EIgl D3 zh dc db o
GLsc tb s
 REC: aaD1 AIgl sh u dl pc pc pb i ng dc   db ir s r ax m uh l axD1
axrD2 GLw ax    v w ax dh ax f h y uuf bc bc bb eD1 EIgl D3 zh dc db o
GLsc    s
sc049.auto:  70.59(60.78)  [H=36, D=2, S=13, I=5, N=51]
Aligned transcription: sc049.hand vs sc049.auto
 LAB: m AIgl    kc kb aa      w ax z s utc Tb     axD1 axrD2 l ax n w
aa l ax w ax zh sh o pc pb i  ng i  n kc kb e n z i ng    tb ax n h
aaD1 AIgl D3 s Trb r rii tc tb
 REC: l AIgl tc gc gb aa l ax w ax z s utc Tb lax w    ax    l ax n
o  l ax w ir zh db o pc    bc bb ii n kc kb e n z i ng tc tb ax n h
aaD1 AIgl ir s trb r rii tc tb
sc050.auto:  63.83(44.68)  [H=30, D=4, S=13, I=9, N=47]
Aligned transcription: sc050.hand vs sc050.auto
 LAB:         chb EIgl    n jhc jhb i ng gc gb iiD1 axD2   h aa f w
EIgl D3   ax    pc pb ax s utc Tb ii          pc pb h i dl kc kb ax m
bb ii kc kb    w  AIgl GLr r ir s ukc Kb ii
 REC: jhc jhb eD1 EIgl D3 n     chb i ng dc db iiD1 axD2 a h aa f
ooD1 OIgl ax bc pc pb ax s dc  db ii kc dc db pc pb   i dl kc kb ax m
```

```
   ii kc kb pc pb AIgl D3  r i  s kc  kb ii
sc051.auto:  68.75(43.75)  [H=33, D=2, S=13, I=12, N=48]
Aligned transcription: sc051.hand vs sc051.auto
 LAB:      AIgl D3 oo  w ir z      nsyl jhb ooD1 OIgl    ir pc pb aaD1
AIgl D3 n tb ax v   l aa    gc gb     ax  w ax n AIgl kc kb ax m
o  f dh ax s kc  kb w      o  sh kc  kb        oo   GLsc tb
 REC: aaD1 AIgl D3 lax w ir z ax m jhc  jhb ooD1 OIgl D3 gc pc pb aaD1
AIgl    n tb ax dnc l aa bc gc gb uuf lax w ax n AIgl kc kb uh m w ax
dl f    ax s ukc kb w oo n kb s  ukc kb w ooD1 OIgl tc   tb
sc052.auto:  75.00(44.23)  [H=39, D=2, S=11, I=16, N=52]
Aligned transcription: sc052.hand vs sc052.auto
 LAB:    jhb ii n    m aaD1 AIgl D3 GLsc prb r       pc pb    eHD1 axD2
m oo   s a m   ax   n   ax n kc kb y uuf    kc kb uh m    bc bb ax   s
    a    m w ax   jhc  jhb ir z         ir f  w ax l uh    kc kb   ii
 REC: sh y   ii n ax m aaD1 AIgl m  pc   prb r ax bc pc pb ii eHD1 axD2
m oo z s   aD1 axD2 n l ax n kc kb y uuf pc kc kb uh m pc bc bb ax z s
aD1 axD2 m   AIgl djhc jhb i  z ax v dc tc tb w ax l uh tb kc kb h ii
sc053.auto:  68.42(52.63)  [H=26, D=4, S=8, I=6, N=38]
Aligned transcription: sc053.hand vs sc053.auto
 LAB: dh lax dl bc bb ii    bc bb i  gc tc trb r uh bnc dl     ir f   y
uuf    dc db eHD1 axD2 tc tb ir tc tb        uh  GLchb dh a  GLs lax f
ir s
 REC: dh lax dl bc    ii dc bc bb ii kc tc trb   uh bnc dl dnc ir f h y
uuf gc dc db           i  tb ax tc tb aa dnc chb utc   Tb ax GLs lax f
ir s
sc054.auto:  68.85(52.46)  [H=42, D=5, S=14, I=10, N=61]
Aligned transcription: sc054.hand vs sc054.auto
 LAB: tb o m   s    e  z  dh ax GLeD1 EIgl    n      GLchb ax n GLsc tb
s aa bc bb z ax f aa m oo s utc Tb aaD1 AIgl    l i sh dh ax m bc bb r
           ir  tb  sh lo  l EIgl D3 l     ax n trc trb r aaD1 AIgl D3
ax   m NO f  s
 REC: tb o m z s lo ir zh db ax GLeD1 EIgl D3 n GLsc chb   ax n
s aa bc    z ax f aa m oo s utc Tb aaD1 AIgl D3 l   ir tb ax m bc bb r
```

```
ax gjhc jhb dnc chb kc klb l EIgl    l i z ax n trc trb r aaD1 AIgl aD1
axD2 m pc pb s
sc055.auto:  59.65(50.88)  [H=34, D=5, S=18, I=5, N=57]
Aligned transcription: sc055.hand vs sc055.auto
 LAB: jhc jhb i dl tc tb uh m bb dl dc db aa f utc Tb ax jhc jhb
      a    kc kb bc   bb ax kc kb ir zh  sh ii w ir zh  sh uh v    dc
db bc bb aaD1 AIgl l i  tc tb   dl m ir s m uh f ax GLsc tb
 REC: jhc jhb i dl tc tb uh m    oo dc db aa f tc  tb ax jhc jhb axD1
axrD2 a gc gb i  GLsc bb ax    gc gb uuf sh         uuf sh aa v ax dc
db pc pb aaD1 AIgl l ir tc tb l ax m ax s m aa f ax tc   tb
sc056.auto:  63.24(47.06)  [H=43, D=4, S=21, I=11, N=68]
Aligned transcription: sc056.hand vs sc056.auto
 LAB:         db o  GLsc tb ax f i l i GLsc pb s r EIgl D3    z  ir n
uh m bb r ax v m    uuf GLsc pb ooD1 OIgl   D3 n GLsc s   ax bc bb a dh
ax pc pb r  f e s  ax z aa tc tb ir kc kb    dl ax n dh ax r rii
s nsyl jhc jhb      lax   n    dl
 REC: dh ax w ax pc tc   tb ax f i l i      bc s r EIgl D3 zh db ir n
uh m ax r ax   m ax kc  pc   pb ooD1 OIgl l ii n      GLs ax m  ax n dh
ax pc pb ax f e tb ax z uh tc tb ax kc kb oo dl o  n    ax r rii tb ax
n m    jhc jhb axD1 axrD2 n ax dl
sc057.auto:  79.49(64.10)  [H=31, D=2, S=6, I=6, N=39]
Aligned transcription: sc057.hand vs sc057.auto
 LAB: dh i  s    n y  uuf dc db    i s    pc plb l     EIgl D3 ir trc
trb r a    GLs m oo    kc kb uh s utc Tb ax m ax z dh ax n e v ax
 REC: dh ir s tc n ii uuf dc db ir z s pc pc plb l eD1 EIgl D3 ir trc
trb   a gc s   m oo tc kc kb aa s utc Tb ax m ax z    ax n e v lax
sc058.auto:  67.24(53.45)  [H=39, D=2, S=17, I=8, N=58]
Aligned transcription: sc058.hand vs sc058.auto
 LAB: m i z ir z s chc chb     uum ax GLsc plb l ii z dc    db ax r ax
n tc tb aaD1 AIgl ax f          a m ax l ii w oe n sh   ii    pc pb  r
jhc jhb uuf s tc m o  dl GLsc       tlb l axD1 uumD2 f ax    tc tb ii
 REC: m i z ir   s chc chb uuf w   ax GLsc plb l ii i dc tc tb ax r i
n tc tb aaD1 AIgl D3 f aD1 axD2 n w ax l ii w e  n GLsh ii bc tc trb ax
```

```
h   y   uuf s tc m ax dl GLsc pb kc klb l      lax   f ax tc tc tb ii
sc059.auto:  63.83(46.81)  [H=30, D=3, S=14, I=8, N=47]
Aligned transcription: sc059.hand vs sc059.auto
 LAB: pb ii tc kc kb u dc nsyl GLsc bb eHD1 tc tb ax sh axD1 axrD2 ax s
i  z s ukc Kb              aa s axD1 uumD2   s uuf   n aa f Tb ax dh
ii a   kc kb s    ax dnc ax    n  tc  tb
 REC: bb ii pc kc kb      ax   n    m  i    tc tb ir sh axD1 axrD2 z  s
ir z s ukc Kb uum aaD1 AIgl D3 s axD1 uumD2 z s uuf e n aa f    ax dh
ii a n kc kb s ir zh db  ax NO th utc tb
sc060.auto:  78.26(60.87)  [H=36, D=1, S=9, I=8, N=46]
Aligned transcription: sc060.hand vs sc060.auto
 LAB:   jhc jhb ooD1 OIgl D3 f i bc dc db ir bc bb a dh ax m      uh
n   ii m i s i ng f ax m dh ax s ii    kc krb r    GLsc    kc kb a sh
      bc bb o kc   kb s
 REC: i jhc jhb ax   w    i  f i gc dc db ax bc bb a dh ax m aaD1 AIgl
n i ng m i s i ng f ax   n  ax s ii tc kc krb r ax GLsc db kc kb a sh
ax tc bc bb o GLsc kb s
sc061.auto:  73.47(53.06)  [H=36, D=1, S=12, I=10, N=49]
Aligned transcription: sc061.hand vs sc061.auto
 LAB: dh ax h           lax s w    l ax r aaD1 AIgl    v ax dh ax m oo
gc gb bc bb ax tc tb w ii n   ax kc kb w oo    tb  r ax n    h aa f
pc pb   aa   s tc tb w oe dl       v
 REC: dh ax h axD1 axrD2 z   s w ax l ax r aaD1 AIgl D3 v ax dh ax m oo
   bc bc bb ax dc db w i  n i kc kc kb w oo sh trb r ax n ax h aa f bc
bb aaD1 AIgl s tc tb w e  dl f utc tb
sc062.auto:  72.92(68.75)  [H=35, D=6, S=7, I=2, N=48]
Aligned transcription: sc062.hand vs sc062.auto
 LAB: w oo tc tb ax w ax z kc kb     a    s ukc Kb eD1 EIgl D3 dnc i ng
db aD1 axD2 n dh ax   m aD1 axD2 n tc tb ir n ax tc tb ax r EIgl D3 tb
ir v n o  GLsc s
 REC: w oo tc tb ax w ax z kc kb aD1 axD2 s ukc Kb                i ng
   aD1 axD2   m  ax n m aD1 axD2 n tc tb ir n ax dc db ax r EIgl D3 tb
ir f n aa NO   s
```

```
sc063.auto:  61.33(48.00)  [H=46, D=5, S=24, I=10, N=75]
Aligned transcription: sc063.hand vs sc063.auto
 LAB:          klb l      aa   r ax w    oe n  GLsc th  r uum ax  f
EIgl D3 z w oe n sh    ii  oo w EIgl D3 z s lax v dc   db h      uh
ng gc gb eHD1 axD2 r riiD1 axD2 n gc gb     uum l   a         sh f o l
axD1 axrD2 dc bc bb aaD1 AIgl D3 r uum bc bb aa bc bb kc krb r uh m bc
bb dl
 REC: dc db pc plb l aaD1 AIgl r ax w ax dc db trc  trb r     lax f eD1
EIgl D3 z w ax n GLchb uuf o  w      i  z s lax   GLsc tb h aaD1 AIgl
n  gc gb eHD1 axD2 r rii   i    n gc gb uuf uum GL1 a GLsc kb sh f o l
          ax bc bb aaD1 AIgl D3 r ax  bc bb aa pc pb kc krb r aa m bc
bb dl
sc064.auto:  70.21(65.96)  [H=33, D=2, S=12, I=2, N=47]
Aligned transcription: sc064.hand vs sc064.auto
 LAB:    w ii  dc bc bb ii h aa dc pc pb   u  sh tc tc  tb ir kc kb a
GLchb dh ax bc bb uh s tc tb ax n y  uuf   kc kb a  s dl tc tb ax n
aaD1 AIgl D3 tc tb
 REC: ax r rii dc bc bb ii h aa tc pc pb w ax sh    utc Tb ax kc kb a
GLchb dh ax bc bb aa s tc tb ax n ii GLuuf kc kb aa s oo tc tb ax n
aaD1 AIgl    tc tb
sc065.auto:  78.95(65.79)  [H=30, D=1, S=7, I=5, N=38]
Aligned transcription: sc065.hand vs sc065.auto
 LAB:    aD1      axrD2 bc bb      u GLsc chb ax m EIgl D3 kc kb s ir z
axD1 axrD2 n pc pb oo kc kb ax n bc bb ii f    s o  s ax jhc jhb i  z

 REC: pb aD1 axD2 m     bc bb ax w i GLsc chb ax m EIgl    kc kb s ir z
axD1 axrD2 n pc pb oo kc kb ax n bc bb ii f ax s aa s ax chc chb ir GLs

sc066.auto:  75.41(65.57)  [H=46, D=4, S=11, I=6, N=61]
Aligned transcription: sc066.hand vs sc066.auto
 LAB: m aa tc tb ir n ax n kc krb r     EIgl D3 gc gc grb r axD1 axrD2
dc db w oo    f chc chb     uum l   ir GLsc pb s ax n   e  gc gb    z
ir bnc ir GLsc dh ax m      oo   l axD1 uumD2 v   ax dh  ax kc kb aD1
```

```
axrD2 n tc tb ii
 REC: m aa    tb      ax n kc krb r eD1 EIgl D3    gc gb  r axD1 axrD2
dc db w oo dl f chc chb uuf uum GLl ir GLsc pb s ax n i ng gc gb ir z
i  v   ax GLsc db ax m ooD1 OIgl l axD1 uumD2 GLw ax dnc ax kc kb aD1
axD2  n tc tb ii
sc067.auto:  77.08(64.58)  [H=37, D=1, S=10, I=6, N=48]
Aligned transcription: sc067.hand vs sc067.auto
 LAB:   aa tc tb uuf kc kb i dc db   z ir dc db          oo dnc dh ax
l ir tc tlb l aaD1 AIgl    s bc bb uh n z w   ax th dh ir chc  chb e
r rii z o n tc tb   o GLsc pb
 REC: a pc tc tb uuf kc kb i    GLsc s ir dc db ooD1 OIgl l  ax  dh ax
l ir tc tlb l aaD1 AIgl D3 s bc bb uh n z w i pc tc Tb ir GLsc chb lax
r rii z o n tc tb h o GLsc pb
sc068.auto:  81.40(72.09)  [H=35, D=3, S=5, I=4, N=43]
Aligned transcription: sc068.hand vs sc068.auto
 LAB: h ii    r    m     e    m bb ax dc db ii n ii dnc ir dnc ax    pc
pb aa s pc pb oo GLsc tb ax gc gb e tflap ax v ii z ax s utc Tb a m
GLsc pb
 REC: h ii ax r ax m aD1 axD2 m    ax dc db ii n        i  dnc ax bc pc
pb aa s pc pb oo GLsc tb ax gc gb e dnc   ax v ii z ax s utc Tb a m
tc   tb
sc069.auto:  77.05(67.21)  [H=47, D=5, S=9, I=6, N=61]
Aligned transcription: sc069.hand vs sc069.auto
 LAB: h ir z ax tc tb e m pc pb s tc  tb ir y uuf z ir z kc krb r e dnc
ir GLsc kb aa dc db f      EIgl D3    dl    dc bc bb ax kc kb ir z ax z
ir    kc kb aD1 axD2 n GLw w ax z ir n ax r     riiD1 axD2 z
 REC: h ir z ax tc tb e         n GLs tb ii y ir  z ir z kc krb r e dnc
i  GLsc kb aa dc db f lo l EIgl D3 ax dl ax bc bc bb ax kc kb ir z ir z
ir tc kc kb aD1 axD2       w ax z ir n ax r rii iiD1  axD2 s
sc070.auto:  68.00(58.00)  [H=34, D=2, S=14, I=5, N=50]
Aligned transcription: sc070.hand vs sc070.auto
 LAB: dh EIgl w ax s utc Tb i   dl f r i s tc tb bc bb AIgl kc kb uh s
utc Tb ax m    z dh ax  dh EIgl D3  ax dc n uh          th i ng tc tb
```

```
ax dc db i kc klb l      e    axD2
 REC: f  ir   w ax s dc  db uuf o  f r i s tc th bc bb AIgl kc kb uh s
tc  tb ax m ax z    lax dh      uuf e  dc n uh s utc tb h  i ng tc tb
ax dc db i kc klb l axD1 eHD1 axD2
sc071.auto:  76.74(62.79)  [H=33, D=3, S=7, I=6, N=43]
Aligned transcription: sc071.hand vs sc071.auto
 LAB: dh ax n lax s tc tb e n db ir dc dh ax f iiD1 irD2 s lo l u    kc
kb i ng    w            uum n db o n   pc pb oo dl z s ukc Kb     a
dl GLsc pb
 REC: dh ax n ax  s tc tb e n db i  v  dh ax f      i    s lo   u tc kc
kb i ng ax w axD1 uumD2 ax  n db o n m pc pb oo dl   s ukc Kb aD1 axD2
dl GLsc tb
sc072.auto:  71.70(52.83)  [H=38, D=4, S=11, I=10, N=53]
Aligned transcription: sc072.hand vs sc072.auto
 LAB: e v     r ax w  uh n tc tb     oo kc kb s  ax v dh ax    bc bb
lax dc db z ax n dh ax    bc bb ii z       bc bb ax dh EIgl D3 n e v
ax  m e n    sh ax n w     o    s pc  pb s
 REC: e v uuf w ax dl uh n tc tb l u dl pc tc tb ax   dh ax bc bc bb
lax       z ir n dh ax dc bc bb ii z ir dc bc bb ax    GLdh i  n e v
lax m e n dc sh ax n w uum o th s upc pb s
sc073.auto:  68.29(56.10)  [H=28, D=3, S=10, I=5, N=41]
Aligned transcription: sc073.hand vs sc073.auto
 LAB: m axD1 uumD2 s  Tb ax dh ax s ii n z w ax f ir dl m db o n
  l axD1 axrD2 kc   kb      eD1 EIgl D3 sh ir n ax n dh ii   a  dl pc
pb s
 REC: m axD1 uumD2 sh Tb ax dh ax s ii n z w ax f i  dl      o n db aa
m l axD1 uumD2 GLsc kb i tb eD1 EIgl D3 sh ir n ir n    EIgl D3 o  GLsc
tb s
sc074.auto:  75.51(71.43)  [H=37, D=0, S=12, I=2, N=49]
Aligned transcription: sc074.hand vs sc074.auto
 LAB: m ir s utc Tb ax v lax n ax n h oe dl pc pb tb msyl s oe dl f tc
tb ax dnc   ir z lax GLsc tb ax n ax l aa     jhb kc kb uh pc pb ax v
kc kb o f   ii
```

REC: m ir s utc Tb ax v lax n ax n h oe dl bc dc db nsyl s uh dl f tc tb ax tflap ir z lax tc tb ax n ax l aa chc chb kc kb uh bc bb ax f kc kb o f i y

sc075.auto: 59.62(50.00) [H=31, D=7, S=14, I=5, N=52]

Aligned transcription: sc075.hand vs sc075.auto

LAB: i GLw w ax z ir sh iiD1 axD2 f lo l uuf kc kb dh ax tflap AIgl D3 bc bb uh m tc tb i n tc tb ax dh ax s EIgl D3 m chc chb a GLsc pb y e s Tb ax dc db eD1 EIgl D3

REC: uuf w ax z ax sh ii ax f lo uum gc gb th ax bc bb uh m kc kb dh i n tb ax dh ax s eD1 EIgl D3 n chc chb aD1 axD2 bc i h y ii e s Tb ax dc db EIgl D3

sc076.auto: 60.71(53.57) [H=34, D=3, S=19, I=4, N=56]

Aligned transcription: sc076.hand vs sc076.auto

LAB: jhc jhb u r i ng ir z l aa s utc Tb y iiD1 axD2 r ax GLy ir n ax v lax s ax tc tb ii dh eD1 EIgl sh e axD2 dc m e n ii h ax l lax r riiD1 axD2 s m axD1 axrD2 m ax n GLsc s

REC: jhc jhb u r rii ng ax z l aa s tc tb ii y axD1 axrD2 GLr rii y uuf n ax v lax z ir z s y ii bc bb EIgl sh i pc m e n ii tb ax l e r riiD1 axD2 s m aD1 axD2 m ax n GLsc s

sc077.auto: 75.00(58.93) [H=42, D=2, S=12, I=9, N=56]

Aligned transcription: sc077.hand vs sc077.auto

LAB: aaD1 AIgl D3 l uh v dc tc tb ax pc pb e dnc dl m AIgl trc trb aaD1 AIgl D3 s ax kc kb dl ax n dh ax bc bb a GLsc kb y aa dc db w ax n ax w ax z ax chc chb aaD1 AIgl D3 dl dc db

REC: aaD1 AIgl l uh f tc tb ax pc tc Tb e dnc l ax m aaD1 AIgl D3 tc trb r aaD1 AIgl D3 s ir kc kb w lax dl ax n dh ax bc bb a gjhc jhb ii aa dc db w ax n aa w ax z ax chc chb ax dl aaD1 AIgl D3 o GLsc tb

sc078.auto: 59.57(40.43) [H=28, D=3, S=16, I=9, N=47]

Aligned transcription: sc078.hand vs sc078.auto

LAB: s uh m pc pb ii pc pb dl f aaD1 AIgl D3 n dh ax w oe dl sh lo l a ng gb w ax jhc jhb v e r rii dc db i f ax kc kb

```
dl tc tc   tb      l     lax  n
 REC: ax s aa m pc pb ii pc bc bb oo dl f aaD1 AIgl    n    ax w lax dl
u  tb l AIgl D3 n  kb w i  dnc jhb ax v ax r rii gc dc db i v ax bc bb
dl    GLsc tb l ax l aD1 axD2 n
sc079.auto:  82.54(74.60)  [H=52, D=3, S=8, I=5, N=63]
Aligned transcription: sc079.hand vs sc079.auto
 LAB: dh ax f uuf dc   db v e   axD2 r rii z f ax m pc plb   EIgl D3 s
tb ax    pc plb l EIgl D3 s    bc bb ax GLsc dh  ax    pc prb r aaD1
AIgl D3 s r    m EIgl D3 n z f eHD1 axD2 l ii kc kb o n s utc Tb ax n
tc   tb
 REC: dh ax f uuf GLsc db v uuf ax   r rii z f ax m pc plb l EIgl D3 s
Tb ax pc pc plb l EIgl D3 s pc bc bb ax      GLn ax tc pc prb r aaD1
AIgl D3 s r ax m EIgl D3 n z f      ir   l ii kc kb o n s     tb ax n
GLsc tb
sc080.auto:  72.22(50.00)  [H=26, D=1, S=9, I=8, N=36]
Aligned transcription: sc080.hand vs sc080.auto
 LAB: dh ax s e  r ax m ax n ii ax v   ax w oe dl m     ii ax n ax w ax
z    m uuf v dc tc    tb ax tc tb                  iiD1 axD2 z
 REC: dh ax s ax r ax m ax n ii ax GLw ax w oe dl m GLm ii a  n ax w ax
z db m uuf   pc tc Tb ir tc tc tb ii ir GLsc db th pc   pb   jhc
sc081.auto:  65.22(41.30)  [H=30, D=3, S=13, I=11, N=46]
Aligned transcription: sc081.hand vs sc081.auto
 LAB:   dh ii   e  r riiD1 axD2 w ax z s axD1 axrD2 th   ax r    l ii
   bc bb l i GLsc s   dh ax GLsc n aD1 axrD2 v            lax GLchb lo
  ii n uh th i ng   r ax m      EIgl D3 n z
 REC: h y  ii y ax r rii   ax   w ax   s axD1 axrD2 th s ax r ax l ii
bc bc bb l i      GLs f  ax      n aD1 axD2  v axD1 uumD2 zh  sh    lo
l ii n uh tb i ng n r ax m ax n i    ng n GLs
sc082.auto:  40.91(22.73)  [H=18, D=3, S=23, I=8, N=44]
Aligned transcription: sc082.hand vs sc082.auto
 LAB:    bb o  bc m        ir dl GLsc kb dh   ax kc kb aD1        axrD2
z aa f utc Tb ax h ii dc db gc gb a dh ax dc dh ax chc  chb   i  kc kb
i n z      e  gc gb z
```

```
 REC: bc bb oo tc m ax l o dc db bc   bb EIgl tc tc tb aD1 axD2 l ax
s aa v tc  tb ax h ii dc db gc gb a          dh ax dchc chb y ii gc gb
i n z EIgl D3 kc kb s
sc083.auto:  74.29(65.71)  [H=52, D=2, S=16, I=6, N=70]
Aligned transcription: sc083.hand vs sc083.auto
 LAB: w ii r rii l ii    w ax dl n ii dc tc tb ir dc db ir f r o  s tc
dh ax f r i   jhc  jhb n aD1 axrD2 dh ax   dh ax s Trb r oo    bc bb r
rii    z      ax v   GLuuf z  dc db aD1 axrD2 tc tb oo l axD1 uumD2 v
ax dh ax    pc plb l EIgl D3 s
 REC: w ii r rii l ii ax w i  dl n ii dc tc tb ir dc db ax f r aa s tc
dh ax f r rii GLsc chb n aD1          axD2 dh ax s trb r oo dc bc bb r
rii zh z ax f ax dnc uuf   zh tc tb aD1 axD2  tc tb oo l axD1 axrD2 v
ax dh ax pc pc plb l AIgl D3 s
sc084.auto:  68.75(64.06)  [H=44, D=4, S=16, I=3, N=64]
Aligned transcription: sc084.hand vs sc084.auto
 LAB: dh ax kc   kb axD1 uumD2    GLsc chb s w lax v    dc    tc tb
dh  ax l e f utc Tb ir n  ax n ax tc tb e m GLsc pb tc tb uum ax v ooD1
OIgl D3 dc dh ax    bc bb l aaD1 AIgl n db pc pb ax dc db e    s trc
Trb r riiD1 axD2 n
 REC:    ax GLsc kb axD1 uumD2 dc db   sh  s w     axD1 uumD2 tc tb tc
dhR ax l e f        tc tb ax n ax tc tb i m kc   kb tc tb uum ax v ooD1
OIgl D3 dc db ax bc bc bb l aaD1 AIgl n db pc pb ax dc db eHD1 s tc
Trb r rii   ax   n
sc085.auto:  61.36(50.00)  [H=27, D=1, S=16, I=5, N=44]
Aligned transcription: sc085.hand vs sc085.auto
 LAB: aa f utc Tb ax dh ax r axD1 uumD2    dc db v iiD1 irD2 z tc tb ax
dh ax s aD1 axD2    th y dl s   ii ir   jhc jhb      uh    ng kc   kb
sh ax n ax h e    dc db
 REC: uh f dc  db ax dh ax r axD1 uumD2 dc dc db   dh   i    s tc tb ax
dh ax s aD1 axD2 tb ir w ax s y ii eHD1 jhc jhb axD1 axrD2 n  GLsc db
sh ax n uh h eHD1 dc db ax
sc086.auto:  74.47(63.83)  [H=35, D=1, S=11, I=5, N=47]
Aligned transcription: sc086.hand vs sc086.auto
```

```
 LAB: h ax r rii v n i  ng gc gb aD1 axD2  n w ax z ax trc trb r aaD1
AIgl    f    dl tc tb uuf    gc gb eHD1 axD2 r ir sh    f ax dh ii ir
kc kb eD1 EIgl D3 zh ax       n
 REC:   ax r rii f n ii n  bc z  aD1 axrD2 n l ax z ax trc trb r aaD1
AIgl D3 f oo tc tc tb uuf bc gc gb eHD1 axD2 r ir GLchb f ax dh ii i
kc kb eD1 EIgl D3 zh ax v utc tb
sc087.auto:  70.45(56.82)  [H=31, D=3, S=10, I=6, N=44]
Aligned transcription: sc087.hand vs sc087.auto
 LAB: h ii gc    gb l i m pc pb s tc dh  ax trc trb r a f i  kc kb w
     oo   dc nsyl    aD1 axD2 tc tb ax v dh  ax    kc kb oo n ax  v
ir z aaD1 AIgl D3
 REC: h i  gc dc db l i m    kc s tc dhR ax trc trb r a f ir kc kb w
ooD1 OIgl v  nsyl dh aD1 axD2    tb ax   dnc ax kc kc kb oo n lax bnc
ir z aaD1 AIgl D3 h ii
sc088.auto:  60.98(47.56)  [H=50, D=5, S=27, I=11, N=82]
Aligned transcription: sc088.hand vs sc088.auto
 LAB: sh  ii h ax dc s   ukc Kb eHD1 axD2 s lo ii dc db aaD1 AIgl D3 v
uh dl z dh ax s    ukc Kb  a     n db   dl bc bb ax f oo        ir
GLw w ax z s pc plb l a tb ax dc db axD1 uumD2 v ax   dh ax f r uh n
      GLsc pb EIgl D3      jhb ir z ax v dh ax tc tb          a  bc bb
l ooD1 OIgl D3 dc   z
 REC: chb ii      ir GLs ukc Kb eHD1 axD2 s Tb ii dc db aaD1 AIgl D3 v
db o  z dh ax s dc db  aD1 axrD2 n db o dc bc bb ax f oo w axD1 axrD2
GLw   ax   s tc tlb l a tb ax dc db      lax   w ax z db ax f r ax n tc
bc bb ir   tb EIgl D3 gjhc jhb ir z ax v ir dc tc tb aD1 axD2 bc bc bb
l oo   h    ir GLsc s
sc089.auto:  62.50(53.12)  [H=20, D=4, S=8, I=3, N=32]
Aligned transcription: sc089.hand vs sc089.auto
 LAB: dh ax kc kb w oe sh chb   ax n    eHD1 axD2 w ax zh sh oo      tb
ax n tc tb ax dh ax    pc pb ooD1 OIgl D3 NO tb
 REC:    ax kc kb w    ax GLchb ax n ii eHD1 axD2 w    ir sh oo GLsc tb
ax n    db ax dh ax pc pc pb oo   n    e  tc tb
sc090.auto:  64.44(48.89)  [H=29, D=2, S=14, I=7, N=45]
```

```
Aligned transcription: sc090.hand vs sc090.auto
 LAB: w    uuf ax    pc plb   uh n zh db  ir n tb ax dc db    aa kc kb
n ax s ax z dh ax kc klb l aD1 axrD2 dc   z ir n  gc gb uh dl f tc
dh  ax m uum n
 REC: w ii w   ax bc pc plb l uh n z  dnc ir n db ax dc db uh gc gb s
n e  s ax   dh ax kc klb l aD1       axD2 z ir ng gc gb ax dl f TDHS
dhR ax m ax  n db gc s
sc093.auto:  73.47(63.27)  [H=36, D=2, S=11, I=5, N=49]
Aligned transcription: sc093.hand vs sc093.auto
 LAB: sh ii  r   i  tc tb lax n    f ax m h o  dl ax dc db EIgl D3
bc bb r o n z dc db bc bb aaD1 AIgl D3 dh ax m e dnc    i  trc trb    r
    EIgl n iiD1 axD2 n s uh n
 REC: sh uuf GLr ax tc tb ax  n tc f ax n h oo dl ir dc db EIgl D3 dc
bc bb r o n z dc db bc bb aaD1 AIgl    dh ax m e dnc ir tb s   trb ax r
eD1 EIgl   ng   i    n s uh n
sc094.auto:  61.70(48.94)  [H=29, D=1, S=17, I=6, N=47]
Aligned transcription: sc094.hand vs sc094.auto
 LAB: dh EIgl D3 s ir jhc jhb e s utc Tb ir dh  ax GLsc dh ii aaD1 AIgl
D3 s w ax           dc  th oo     ax z s uuf     n ax z s    pc pb  r i
ng ax r aaD1 AIgl D3 v  dc db
 REC: dh      i  s ir z   db  i s dc  db ir dnc ir dc   db ii aaD1 AIgl
D3 s w ax GLsc pb f utc Tb oo GLw ax z s uuf uum n ax z s pc pc prb r i
ng ax r aaD1 AIgl h  ir tc tb
sc095.auto:  73.33(53.33)  [H=33, D=3, S=9, I=9, N=45]
Aligned transcription: sc095.hand vs sc095.auto
 LAB:          gb oo      dc nsyl z w            lax dc z w ax l o    s
utc Tb ax m i dnc s tc  dh ax  h uh bc bb uh bnc ax v dh ax kc krb
aD1 axD2 dnc   ir dc db h oo dl
 REC: pc pb kc kb oo tc n ax n    z w axD1 axrD2 lax    z w ax l o tb s
utc Tb ax m i     s utc Tb lax h uh bc bb uh v   ax v dh ax kc krb r
aD1 axD2 tflap ir    z  h oo dl
sc096.auto:  70.45(63.64)  [H=31, D=0, S=13, I=3, N=44]
Aligned transcription: sc096.hand vs sc096.auto
```

```
 LAB: m i  s utc Tb ax f oo s aaD1 AIgl D3    th h ax dchc chb iiD1
axD2 dc    dh ax tc tb ii   m o n aD1 axD2 f ax f aaD1 AIgl D3 v h
axrD2 dl y   iiD1 axD2 z
 REC: m ir s utc Tb ax f oo s aaD1 AIgl D3 tc tb h ir dchc chb y
i    dc db ir s  tc tb ii n m o n aD1 axD2 f ax f aaD1 AIgl e  v h
oo    dl dnc ii   i    s
sc097.auto:  68.42(50.00)  [H=26, D=2, S=10, I=7, N=38]
Aligned transcription: sc097.hand vs sc097.auto
 LAB: i GLsc s ii   m  z ir f s    uuf    z ax n db        uh z    oo
dl dh ax chc chb oo z f ax dh ir s h aD1 axD2 s h ax dl      dc db
 REC: i GLsc s ii n ir z ir f s lo uuf ax z ax n db lax dl ax z db oo
dl v  ax chc chb oo z f ax dh i  z h          a z db dl GLsc pb chc
sc099.auto:  70.59(45.10)  [H=36, D=2, S=13, I=13, N=51]
Aligned transcription: sc099.hand vs sc099.auto
 LAB:      bc bb e      th chc chb eHD1 axD2    dc dh ax f lax   s m ii
tc tb i ng ir n oo dnc ax tc tb uum axD1 axrD2 bc bb l aaD1 AIgl D3 jhc
jhb          h  lax  GL axD1 uumD2   v ax w    lax GLsc bb      o  s

 REC: m ax bc bb e h ax th chc chb eHD1 axD2 dc dc db ax f lax z s m ii
tc tb      ii n ax w   ax tc tb uum ax   bc    bc bb l aaD1 AIgl D3 jhc
jhb db bc bb ax GLsc f  axD1 uumD2 u v ax w ax gc  bc   bb aa h ax GLs

sc100.auto:  66.67(48.15)  [H=18, D=1, S=8, I=5, N=27]
Aligned transcription: sc100.hand vs sc100.auto
 LAB: aaD1 AIgl v jhc  jhb ax s ii n chb   aa  dl z   tb e r i   ng o f
i  n ax h           uh r      rii
 REC: aaD1 AIgl   djhc jhb ax s ii n GLchb lax dl GLs tb e r i n db o f
ax n ax h aaD1 AIgl D3 r ir h ii
```

# Appendix C

# TR Classes

| TR Class | Left-hand Phone | Right-hand Phone |
|---|---|---|
| TR1 | vowel, dl, r, w, y or diphthongal glide | stop-closure, no, mo, thS or dhS |
| TR2 | vowel, dl, r, w, y or diphthongal glide | -VOICE fricative or affricate-release or burst, or h |
| TR3 | vowel, dl, r, w, y or diphthongal glide | +VOICE fricative or affricate-release or burst, or tflap, dnc, bnc or gnc |
| TR4 | vowel, dl, r, w, y or diphthongal glide | nasal |
| TR5 | vowel, dl, r, w, y or diphthongal glide | l or cl |
| TR67 | vowel, dl, r, w, y or diphthongal glide | GL, GLxc, GLxyc, GSvowel |
| TR8 | nasal | vowel, y or diphthongal glide |
| TR10 | nasal | r, w, dl, l or nasal |
| TR11 | nasal | +VOICE fricative or affricate-release |
| TR12 | nasal | -VOICE fricative or affricate-release, or h |
| TR13 | nasal | GL, GLxc, GLxyc, GSvowel |
| TR14 | nasal | +VOICE stop-closure |
| TR15 | nasal | -VOICE stop-closure |
| TR16 | SIL, GL, GSvowel, GLxc, GLxyc, NP | vowel, r, w, l, cl, y or diphthongal glide |
| TR17 | -VOICE fricative or affricate-release or burst, or h, lo, or dlo | vowel, dl, l, cl, y or diphthongal glide |
| TR18 | +VOICE fricative or affricate-release or burst, or tflap, dnc, bnc or gnc | vowel, dl, l, cl, y or diphthongal glide |
| TR19 | l, cl or GLl | vowel, y or diphthongal glide |
| TR20a | z, zh, jhb, s, sh or chb | tc, utc, dc, chc, jhc, t*c, d*c |
| TR20b | z, zh, jhb, s, sh or chb | pc, upc, bc, p*c, b*c |
| TR20c | z, zh, jhb, s, sh or chb | kc, ukc, gc, k*c, g*c |

| TR Class | Left-hand Phone | Right-hand Phone |
|---|---|---|
| TR20d1 | f, th, PVdh, v | stop-closure, dhS, thS, GSvowel |
| TR20d2 | z, zh, jhb, s, sh or chb | dhS, thS, GSvowel |
| TR20e | z, zh, jhb, s, sh or chb | trc, drc |
| TR21 | +VOICE fricative or affricate-release | +VOICE fricative or affricate-release or burst, or tflap, dnc, bnc or gnc |
| TR22 | fricative, affricate-release, dnc, bnc, or gnc | -VOICE fricative or affricate-release, or h, trb or Trb |
| TR23 | -VOICE fricative or affricate-release | r or w |
| TR24 | +VOICE fricative or affricate-release, or dnc, bnc or gnc | r or w |
| TR25 | trb, Trb, drb, prb, Prb, brb, krb, Krb, or grb | r or vowel or diphthongal glide |
| TR26 | -VOICE burst | r or w |
| TR27 | +VOICE burst | r or w |
| TR28 | -VOICE fricative or affricate-release or burst | nasal |
| TR29 | +VOICE fricative or affricate-release or burst, or dnc, bnc or gnc | nasal |
| TR30a | SIL, NP, -VOICE stop closure or complex closure with -VOICE final element | nasal, bc, dc, gc or jhc |
| TR30b | nasal, vowel, diphthongal glide, dl, bc, dc, gc or complex closure with +VOICE final element | nasal, GLnasal, or GL-VOICEconsonant |
| TR31 | tc, SIL, dhS, thS, nasal, no | tb, Tb, tlb, dhR, thR |
| TR32 | dc, SIL, nasal | db, dlb |
| TR33 | bc, SIL, nasal | bb, blb |
| TR34 | pc, upc, SIL, nasal, mo | pb, Pb, plb |
| TR35 | gc, SIL, nasal, ngo | gb, glb |
| TR36 | kc, ukc, SIL, nasal, ngo | kb, Kb, klb |

| TR Class | Left-hand Phone | Right-hand Phone |
|---|---|---|
| TR37 | chc, SIL, GL*chc | chb |
| TR38 | jhc, SIL | jhb |
| TR39 | tc, dc, SIL, nasal, trc, drc | trb, Trb, drb |
| TR40 | pc, bc, SIL, nasal | prb, Prb, brb |
| TR41 | kc, gc, SIL, nasal | krb, Krb, grb |
| TR42 | stop-closure, GL*xc except GLchc and GLjhc, SIL, no, mo or ngo | fricative or affricate-release or h |
| TR43 | burst | stop-closure |
| TR44 | -VOICE fricative or affricate-release | +VOICE fricative or affricate-release |
| TR45 | diphthongal glide | vowel, dl, y, r, w, GLr, GLw |
| TR46 | fricative or affricate-release | -VOICE burst |
| TR47 | nasal | PNASdh |
| TR51 | utc | Tb or tb |
| TR52 | fricative or affricate-release | lo or dlo |
| TR53 | burst | fricative |
| TR60 | monophthongal vowel | monophthongal vowel |

# Bibliography


BAHL, L.R., P.F.BROWN, P.V.DE SOUZA, & R.L.MERCER. 1989. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-37 (7).

BARRY, W.J., & A.J. FOURCIN. 1990. Levels of labelling. *Speech, Hearing and Language* 4.29–44.

BATES, S., 1995. *Towards a Definition of Schwa: An Acoustic Investigation of Vowel Reduction in English*. University of Edinburgh dissertation.

BECKMAN, M.E., & J.EDWARDS. 1992. Intonational categories and the articulatory control of duration. In *Speech Perception, Production and Linguistic Structure*, ed. by E. Tokhura & Y.Sagisaka, 356–375. Tokyo OHM Publishing.

BENGIO, Y. 1996. *Neural Networks for Speech and Sequence Recognition*. International Thompson Computer Press.

BISHOP, C.M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

BISIANI, R., T. ANANTHARAMAN, & L. BUTCHER. 1989. BEAM: An accelerator for speech recognition. In *Proceedings of the ICASSP*.

BLUMSTEIN, S.E., & K.N. STEVENS. 1979. Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America* 66.1001–1017.

——, & ——. 1980. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America* 67.648–662.

BOCCHIERI, E.L., & G.R.DODDINGTON. 1986. Frame-specific statistical features for speaker-independent speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-34.

BOURLARD, H. 1995. Towards increasing speech recognition error rates. In *Proceedings of Eurospeech*, 883–894.

BROWMAN, C.P., & L. GOLDSTEIN. 1990. Tiers in atriculatory phonology, with some implications for casual speech. In *Papers in Laboratory Phonology I*, ed. by J.Kingston & M.Beckman, 341–376. Cambridge University Press.

CATFORD, J.C. 1977. *Fundamental Problems in Phonetics*, 112–116. Edinburgh University Press.

CHOW, Y-L., R. SCHWARTZ, & OTHERS. 1986. The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system. In *Proceedings of the ICASSP*, 1593–1596.

CLARK, J., & C. YALLOP. 1990. *An Introduction to Phonetics and Phonology*. Basil Blackwell.

DALBY, J., A.CROWE, & A.SUTHERLAND. 1989. Formant-based vowel classification in continuous speech. In *Proceedings of Eurospeech*, volume 2.

DAVIS, S.B., & P. MERMELSTEIN. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-28 (4).

DUDA, R.O., & P.E. HART. 1973. *Pattern Classification and Scene Analysis*. John Wiley.

FANT, G. 1960. *Acoustic Theory of Speech Production*. Mouton.

—— 1995. The LF-model revisited. Transformations and frequency-domain analysis. *Speech Transmission Laboratory Quarterly Report* .

FLURY, B., & H.RIEDWYL. 1988. *Multivariate Statistics – A Practical Approach*. Chapman and Hall.

FRENCH, N.R., & J.C.STEINBERG. 1947. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America* 19.90–119.

GAROFOLO, J.S., L. LAMEL, & W. FISHER, 1990. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. U.S. Department of Commerce, Gaithersburg, MD 20899.

HADI, A.S. 1996. *Matrix Algebra as a Tool.* Duxbury Press.

HAEB-UMBACH, R., & H.NEY. 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of the ICASSP*, volume 1, 13–16.

HARRINGTON, J. 1987. Automatic recognition of English consonants. In *Aspects of Speech Technology*, ed. by M.Jack & J.Laver. Edinburgh University Press.

HAWKINS, S. 1995. Arguments for a nonsegmental view of speech perception. In *Proceedings of the International Congress of Phonetic Sciences*, volume 3, 18–25.

HUANG, X.D., Y. ARIKI, & M.A. JACK. 1991. *Hidden Markov Models for Speech Recognition.* Edinburgh Information Technology Series. Edinburgh University Press.

HWANG, M.Y., F. ALLEVA, & X. HUANG. 1993. Senones, multi-pass search, and unified stochastic modelling in Sphinx-II. In *Proceedings of Eurospeech*, volume 3, 2143–2146.

KOHONEN, T. 1989. *Self-Organisation and Associative Memory.* Springer Series in Information Sciences. Springer-Verlag.

KONDO, Y., 1995. *Production of Schwa by Japanese Speakers of English: A Cross-Linguistic Study of Coarticulatory Strategies.* University of Edinburgh dissertation.

KUEHN, D.P., & K.L. MOLL. 1976. A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics* 4.303–320.

LAMEL, L.F., & J.L.GAUVAIN. 1993. High performance speaker-independent phone recognition using CDHMM. In *Proceedings of Eurospeech*, volume 1, 121–124.

LEE, C.H., L.R. RABINER, & OTHERS. 1990a. Acoustic modelling for large vocabulary speech recognition. *Computer Speech and Language* 4.127–165.

LEE, K-F., H-W.HON, & R.REDDY. 1980. An overview of the Sphinx recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing* .

——, & H.W. HON. 1988. Large vocabulary speaker-independent continuous speech recognition. In *Proceedings of the ICASSP*.

——, S.HAYAZIMU, H-W.HON, C.HUANG, J.SCHWARTZ, & R.WEIDE. 1990b. Allophone clustering for continuous speech recognition. In *Proceedings of the ICASSP*, 749–752.

LEE, KAI-FU. 1990. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* .

LINDBLOM, B. 1983. Economy of speech gestures. In *The Production of Speech*, ed. by P.MacNeilage. Springer Verlag.

LOCAL, J. 1995. Making sense of dynamic, non-segmental phonetics. In *Proceedings of the International Congress of Phonetic Sciences*, volume 3, 3–9.

MARDIA, K.V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57.519–530.

MCINNES, F.R. 1991. Context-sensitive phoneme lattice generation using interpolated demi-diphone and triphone models. In *Proceedings of Eurospeech*, 639–642.

——, D.MCKELVIE, & S.M.HILLER. 1990. The structure, strategy and performance of a modular continuous speech recognition system. In *Proceedings of the Institute of Acoustics*, volume 12, 173–180.

——, M.A.JACK, & J.LAVER. 1989a. Template adaptation in an isolated word-recognition system. In *Proceedings of IEE*, volume 136, 119–126.

——, Y.ARIKI, & A.A.WRENCH. 1989b. Enhancement and optimisation of a speech recognition front end based on hidden Markov models. In *Proceedings of Eurospeech*, volume 2, 461–464.

NG, K., & V.W.ZUE. 1997. Subword unit representations for spoken document retrieval. In *Proceedings of Eurospeech*.

NOLL, A.M. 1964. Short-time spectrum and "cepstrum" techniques for voice pitch detection. *Journal of the Acoustical Society of America* 36.296–302.

OHALA, J. 1989. Respiratory activity in speech. In *Speech Production and Speech Modelling*, ed. by W.Hardcastle & A.Marchal, NATO ASI. Kluwer.

O'SHAUGNESSEY, D. 1987. *Speech Communication: Human and Machine*. Addison Wesley.

OSHIKA, B.T., V.W. ZUE, & OTHERS. 1975. The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-23 No 1.

OWENS, F.J. (ed.) 1993. *Signal Processing of Speech.* Macmillan New Electronics Series. Macmillan.

PIRANI, G. (ed.) 1990. *Advanced Algorithms and Architectures for Speech Understanding.* Springer Verlag. ESPRIT Research Reports Project 26.

RABINER, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77.257–285.

RUSSELL, M.J., & A.E. COOK. 1987. Experimental evaluation of duration modelling techniques for automatic speech recognition. In *Proceedings of the ICASSP*, 2376–2379.

——, & R.K. MOORE. 1985. Explicit modelling of state-occupancy in hidden Markov models for automatic speech recognition. In *Proceedings of the ICASSP*, 5–8.

SCHAFER, R.W., & L.R. RABINER. 1975. Digital representations of speech signals. In *Proceedings of the IEEE*, volume 63 (4), 662–667.

SCHALKOFF, R. 1992. *Pattern Recognition: Statistical Structural and Neural Approaches.* John Wiley.

SCHARF, BERTRAM. 1972. Critical bands. In *Foundations of Modern Auditory Theory*, ed. by J.V. Tobias, volume 1. Academic Press.

SMITH, D.K. 1991. *Dynamic Programming: A Practical Introduction.* Ellis Horwood Series in Mathematics and its Applications. Ellis Horwood.

VEENEMAN, D.E., & S.L. BEMENT. 1985. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-33 No 2.

WOODLAND, P.C., & S.J.YOUNG. 1993. The HTK tied-state continuous speech recogniser. In *Proceedings of Eurospeech*, volume 3, 2207–2210.

YOUNG, S.J., 1992. The HTK Hidden Markov Model Toolkit V1.4 Reference Manual.

—— 1996. Large vocabulary continuous speech recognition: a review. Technical report, Cambridge University Engineering Department, Speech Group.

——, J.ODELL, & OTHERS. 1997. *The HTK Book.*

——, J. J. ODELL, & P. C. WOODLAND. 1994. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of ARPA Workshop on Human Language Technology.*

——, & P.C.WOODLAND. 1993. The use of state-tying in continuous speech recognition. In *Proceedings of Eurospeech*, volume 3, 2203–2206.

ZUE, V., J.GLASS, M.PHILLIPS, & S.SENEFF. 1989. The MIT SUMMIT speech recognition system: A progress report. In *Proceedings of DARPA Speech and Natural Language Workshop*, 179–189. Morgan Kaufman.

ZWICKER, E. 1961. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Journal of the Acoustical Society of America* 33.248.

——, & E. TERHARDT. 1980. Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America* 68.1523–1525.