AUTOMATIC PROSODIC ANALYSIS FOR COMPUTER AIDED PRONUNCIATION TEACHING

Paul Christopher Bagshaw

A thesis submitted for the degree of
Doctor of Philosophy

1994

Abstract

Correct pronunciation of spoken language requires the appropriate modulation of acoustic characteristics of speech to convey linguistic information at a suprasegmental level. Such prosodic modulation is a key aspect of spoken language and is an important component of foreign language learning, for purposes of both comprehension and intelligibility. Computer aided pronunciation teaching involves automatic analysis of the speech of a non-native talker in order to provide a diagnosis of the learner's performance in comparison with the speech of a native talker. This thesis describes research undertaken to automatically analyse the prosodic aspects of speech for computer aided pronunciation teaching.

It is necessary to describe the suprasegmental composition of a learner's speech in order to characterise significant deviations from a native-like prosody, and to offer some kind of corrective diagnosis. Phonological theories of prosody aim to describe the suprasegmental composition of speech for a specific language. It is argued here that the suprasegmental composition of the speech of a non-native talker can be highly influenced by mother-tongue interference thereby rendering a language-specific phonological representation of prosody inappropriate. Moreover, languages vary in the way acoustic characteristics of speech are modified to manifest prosodic aspects of speech and the only secure means available to describe prosody for foreign language teaching therefore lies in an acoustic-phonetic representation. The automatic prosodic analysis of speech presented in this thesis aims to provide such an acoustic-phonetic representation.

The prosodic aspects of speech are described in a syllabic domain which is synchronised with a phonetic segmentation. An algorithm is presented which groups acoustic-phonetic segments into syllabic units. The acoustic-phonetic syllabification is shown to correlate with phonological syllabification. The fundamental frequency ($F\emptyset$) of speech, the duration and energy of phonetic units and the vowel quality of syllable nuclei play important roles in characterising the prosodic features of stress, rhythm, and intonation. The determination of $F\emptyset$ is required as an initial process in the automatic prosodic analysis of speech. The problems of determining $F\emptyset$ in a way which minimises errors in prosodic analysis are addressed, since the $F\emptyset$ contour of an utterance is affected by segmental content, micro-perturbations, the talker's anatomy and physiology together with errors involved in its determination from the speech waveform. Methods of speaker normalisation and piecewise stylisation of $F\emptyset$ contours are described and a method to process the $F\emptyset$ contour in order to locate and characterise pitch accents and thus provide an acoustic-phonetic description of intonation, is highlighted. Measurements of duration, energy and vowel quality are investigated with respect to their correlation with sentential stress. The process of analysing syllable prominence is complicated by the interaction of these acoustic features in the manifestation of stress and by the fact that they are

also influenced by factors other than stress. The duration, energy and vowel quality of phonetic units vary due to acoustic-phonetic context, syllable length and syllable prominence levels. The research described in this thesis aims to normalise acoustic features for non-prosodic aspects of speech and to combine the processed acoustic features to form a prosodic description of speech. The combination of the acoustic features assumes that stress is predominantly marked by variations in duration, energy and vowel quality, and that pitch accents are marked by the melody of fundamental frequency. *FØ* can be used as a secondary cue to the location of prominent syllables because pitch accents are observed to fall on prominent syllables.

The resultant, automatically determined, prosodic description is shown to be useful in comparing the prosodic aspects of the speech of a non-native learner of English with the speech of a native English talker.

Zusammenfassung

Automatische prosodische Analyse für Computer-unterstützten Ausspracheunterricht

Die angemessene Modulation der akustischen Merkmale der Sprache ist notwendig für die korrekte Aussprache gesprochener Sprache, damit die linguistische Information auf suprasegmentaler Ebene vermittelt werden kann. Eine solche prosodische Modulation ist Schlüsselaspekt der gesprochenen Sprache und ein wichtiger Bestandteil des Lernens von Fremdsprachen, sowohl für das Verstehen als auch für das Verstandenwerden. Computerunterstützter Ausspracheunterricht beinhaltet automatische Sprachanalyse eines Nicht-Muttersprachlers damit eine Diagnose der Leistung des Schülers im Vergleich zu der eines Muttersprachlers gemacht werden kann.

Es ist notwendig, die suprasegmentale Zusammentstellung der Sprache eines Schülers zu beschreiben, um die bedeutenden Abweichungen von der muttersprachlichen Prosodie zu charakterisieren, und um eine Art korrektive Diagnose zu bieten. Phonologische Theorien der Prosodie beabsichtigen die suprasegmentale Zusammenstellung einer bestimmten Sprache zu beschreiben. Es wird hier argumentiert, daß die suprasegmentale Zusammenstellung der Sprache eines Nicht-Muttersprachlers deutlich durch die eigene Muttersprache beeinflußt werden kann und damit eine sprach-spezifische phonologische Representation unangemessen macht. Außerdem varieren Sprachen in der Art in der die akustischen Charakteristiken modifiziert werden, um die prosodischen Aspekte der Sprache zu manifestieren und die einzig sichere Art, um Prosodie für Fremdsprachenunterricht zu beschreiben, ist deshalb die akustisch-phonetische Representation. Die automatische prosodische Sprachanalyse, die in dieser Dissertation vorgestellt wird, versucht eine solch akustisch-phonetische Representation zu bieten.

Die prosodischen Aspekte der Sprache werden in einer Silbendomäne beschrieben, welche mit einer phonetischen Segmentation synchronisiert ist. Ein Algorithmus wird vorgestellt, der akustisch-phonetische Segmente in Silbeneinheiten gruppiert. Es wird gezeigt, daß die akustisch-phonetische Versilbung mit der phonologischen Versilbung korreliert. Die Grundfrequenz ($F\emptyset$) von Sprache, die Dauer und Energie phonetischer Einheiten und die Vokalqualität vom Silbennuklei spielen eine wichtige Rolle in der Charakterisierung der prosodischen Merkmale Rhythmus, Intonation und Betonung. Die Feststellung von $F\emptyset$ ist notwendig als anfänglicher Prozeß in der automatischen prosodischen Sprachanalyse. Die Probleme werden addressiert, die auftreten, wenn man versucht $F\emptyset$ so festzustellen, daß minimale Fehler in der prosodischen Analyse auftreten, da die $F\emptyset$-Kontur einer Äußerung beeinflußt wird durch den segmentalen Inhalt, Mikrostörungen, Anatomie und Physiologie des Sprechers und die Fehler, die zusammenhängen mit

der Determination der Sprachwellenform. Methoden der Sprechernormalisierung und die teilweise Stilisation von *FØ* Kontouren werden beschrieben und eine Methode wird hervorgehoben, um *FØ* Kontouren zu verarbeiten damit die Tonhöhenakzente lokalisiert und charakterisiert werden können und dadurch eine akustisch-phonetische Beschreibunger der Intonation geboten wird. Messungen von Dauer, Energie und Vokalqualität werden untersucht hinsichtlich ihrer Korrelation zur Satzbetonung. Der Prozess der Analyse von Silbenprominenz wird verkompliziert durch die Interaktion dieser akustischen Merkmale in der Manifestation von Betonung und dadurch, daß sie ebenfalls durch Faktoren außer Betonung beeinflußt werden. Die Dauer, Energie und Vokalqualität der phonetischen Einheiten varieren aufgrund des akustisch-phonetischen Kontextes, Silbenlänge und Silbenprominenzhöhen. Die Forschung, die in dieser Arbeit beschrieben wird, versucht die akustischen Merkmale zu normalisieren für non-prosodische Aspekte der Sprache, und die verarbeiteten akustischen Merkmale zu kombinieren, um eine prosodische Beschreibung der Sprache zu bilden. Die Kombination von akustischen Merkmalen nimmt an, daß Betonung überwiegend durch Variationen in Dauer, Energie und Vokalqualität markiert wird, und daß Tonhöhenakzente durch die Melodie der Grundfrequenz gekennzeichnet werden. *FØ* kann als sekundäres Merkmal für die Lokalisierung von prominenten Silben dienen, da festgestellt werden kann, daß die Tonhöhenakzente auf prominente Silben fallen.

Es wird gezeigt, daß die daraus resultierende, automatisch determinierte, prosodische Beschreibung nützlich ist, für den Vergleich von prosodischen Aspekten der Sprache eines nicht-muttersprachlichen Englischschülers mit der Sprache eines englischen Muttersprachlers.

Translated by Miriam Eckert

Résumé

Analyse automatique de la prosodie pour l'enseignement de la prononciation à l'aide d'ordinateurs

La prononciation correcte du langage oral nécessite la modulation appropriée des caractéristiques acoustiques de la parole afin de communiquer l'information linguistique à un niveau suprasegmental. Une telle modulation prosodique est un aspect-clé du langage oral et une composante importante de l'apprentissage d'une langue étrangère, pour des buts tant de compréhension que d'intelligibilité. L'enseignement de la prononciation à l'aide d'ordinateurs nécessite l'analyse automatique de la parole d'une personne dont la langue en question n'est pas la langue maternelle afin de procurer un diagnostic de la performance de l'élève en comparaison avec la parole d'un locuteur natif. Cette thèse décrit la recherche entreprise afin d'analyser automatiquement les aspects prosodiques de la parole pour l'enseignement de la prononciation à l'aide d'ordinateurs.

Il est nécessaire de décrire la composition suprasegmentale de la parole de l'élève afin de caractériser des déviations significatives de la prosodie native, et afin d'offrir un certain diagnostic de correction. Les théories phonologiques de la prosodie ont pour but de décrire la composition suprasegmentale de la parole pour une langue spécifique. L'argument considéré dans ce document est que la composition suprasegmentale de la parole d'un locuteur non-natif peut être grandement influencée par l'interférence de sa langue maternelle propre, ceci rendant alors inappropriée une représentation phonologique de la prosodie spécifique au langage. De plus, les langues varient au niveau de la manière dont les caractéristiques acoustiques de la parole sont modifiées pour manifester les aspects prosodiques de la parole, et le seul moyen sûr disponible pour décrire la prosodie pour l'enseignement d'une langue étrangère se situe donc au niveau d'une représentation acoustico-phonétique. L'analyse prosodique automatique de la parole présentée dans cette thèse a pour but de procurer une telle représentation acoustico-phonétique.

Les aspects prosodiques de la parole sont décrits dans un domaine syllabique qui est synchronisé avec une segmentation phonétique. On présente un algorithme regroupant les segments acoustico-phonétiques en unités syllabiques. Il est démontré que la syllabification acoustico-phonétique est en corrélation avec la syllabification phonologique. La fréquence fondamentale ($F\emptyset$) de la parole, la durée et l'énergie des unités phonétiques et la qualité vocalique des noyaux syllabiques jouent un rôle important dans la caractérisation des traits prosodiques d'accent, rythme et intonation. Il est nécessaire de déterminer $F\emptyset$ initialement pour l'analyse prosodique automatique de la parole. On considère les problèmes posés par la nécessité de déterminer $F\emptyset$ de manière à minimiser

les erreurs d'analyse prosodique, étant donné que le contour de *FØ* d'un énoncé est affecté par le contenu segmental, les micro-pertubations, l'anatomie et la physiologie du locuteur ainsi que les erreurs entraînées par sa détermination basée sur le signal de la parole. On décrit des méthodes de normalisation du locuteur et de stylisation pièce par pièce des contours de *FØ* et on fait apparaître une méthode permettant de traiter les contours *FØ* afin de localiser et caractériser les accents intonatifs et par là de procurer une description acoustico-phonétique de l'intonation. On examine des mesures de durée, énergie, et qualité vocalique du point de vue de leur corrélation avec l'accent de phrase. Le procédé d'analyse de la proéminence de syllabe est compliqué par l'interaction de ces traits acoustiques dans la manifestation de l'accent et par le fait qu'elles sont aussi influencées par des facteurs autres que l'accent. La durée, l'énergie et la qualité vocalique des unités phonétiques varient en fonction du contexte acoustico-phonétique, de la longueur des syllabes, et des niveaux de proéminence des syllabes. La recherche décrite dans cette thèse a pour but de normaliser les traits acoustiques pour les aspects non-prosodiques de la parole, et de combiner les traits acoustiques traités afin de former une description prosodique de la parole. La combinaison de traits acoustiques suppose que l'accent est caractérisé principalement par des variations de durée, énergie et qualité vocalique, et que les accents intonatifs sont caractérisés par la mélodie de la fréquence fondamentale. *FØ* peut être utilisé comme indice secondaire quant à la localisation des syllabes proéminentes étant donné qu'on observe une chute des accents intonatifs sur les syllabes proéminentes.

On démontre que la description prosodique résultante, déterminée automatiquement, est utile à la comparaison des aspects prosodiques de la parole d'une personne dont l'anglais n'est pas la langue maternelle avec la parole d'une personne dont l'anglais est la langue maternelle.

Translated by Nathalie Vergeynst

Acknowledgements

The production of this thesis and the work reported within it have been influenced by colleagues in the fields of engineering, speech technology and linguistics. I am grateful to many people (of which there are too many to remember) who, through formal and more often informal discussions, have moulded some of the fundamental linguistic concepts which are implicit to the research described here; particularly colleagues and friends at the Centre for Speech Technology Research and those who attended the "Elsnet Summer School on Prosody: Integration of Speech and Natural Language through Prosody", held at University College London during July 1993. There are a number of individuals who qualify for special thanks (no order of significance implied):

Prof. Mervyn Jack and Prof. John Laver; for their roles as my official supervisors and for their support and constructive criticisms.

Dr. Alan Wrench; for spending time to proof read the thesis draft manuscript and for pointing out some of its earlier discrepancies and ambiguities.

Dr. Jim Hieronymus, Mr. Stephen Isard, Dr. Bob Ladd, Dr. Nick Campbell, and Prof. Jacqualine Vaissière; for suggestions and discussions related to the research presented here.

Ms. Sally Bates; for her valuable advice on the subject of vowel reduction and the perceptual classification of vowels.

Dr. Eddie Rooney and Dr. Steve Hiller; for their discussions on the subject of the SPELL project.

Dr. Fergus McInnes; for his assistance with some of the mathematics and for introducing me to the concept of entropy.

Dr. Peter Meer, at the State University of New Jersey; for introducing me to the method of least median of squared residuals regression analysis.

Dr. Briony Williams, for her suggestions and assistance regarding morphology and syllabification on a phonological basis.

Dr. John Elliott; for assistance with LaTeX, used in the preparation of this thesis.

Ms. Sue Fitt and Ms. Tina Scott; for enduring my frequent questions about English grammar, usage and spelling.

Dr. Kurematsu; for his kind support of my visit to the ATR Interpreting Telephony Research Laboratories, Kyoto as part of these studies.

Those involved in the implementation of some of the algorithms investigated. The implementation of the Feature-based *FØ* tracker (FBFT) is the work of Bob Brennan and Michael Garris. Both the Cepstral-based *FØ* determinator (CFD) and the Parallel processing method (PP) were implemented by Dr. Steve Hiller.

Dr. Daniel Hirst, at Institut de Phonétique d'Aix-en-Provence; for providing an implementation of the algorithm he developed to model *FØ* contours using a quadratic spline function.

Those anonymous people working at the Science and Engineering Research Council who have conducted the background administrative duties which have financed these research activities.

The computing staff at CSTR; for keeping the tools of research in working order (to the best of their abilities).

Finally, many thanks to Family and Ioana Donescu for the mental support and encouragement.

Contents

Abstract i

Declaration of Originality vii

Acknowledgements viii

List of Figures xiii

List of Tables xv

Glossary of Technical Abbreviations xvi

1 Introduction 1

2 Prosodic Description in Foreign Language Teaching 7
- 2.1 The need for prosody in foreign language teaching . . . 8
  - 2.1.1 Stress . . . 9
  - 2.1.2 Intonation . . . 14
- 2.2 Phonological theories of prosody . . . 16
  - 2.2.1 Degrees of stress . . . 17
  - 2.2.2 Configuration theory . . . 19
  - 2.2.3 Tone sequence theory . . . 22
  - 2.2.4 Accented and stressed syllables: Terminology . . . 27
- 2.3 Discussion . . . 29

3 Acoustic-Prosodic Analysis: An overview of related work 31
- 3.1 Acoustic parameters for prosodic analysis . . . 32
- 3.2 Automatic extraction of fundamental frequency . . . 37
  - 3.2.1 Cepstrum-based *FØ* determinator (CFD) . . . 39
  - 3.2.2 Harmonic product spectrum (HPS) . . . 41
  - 3.2.3 Feature-based *FØ* tracker (FBFT) . . . 44
  - 3.2.4 Parallel processing method (PP) . . . 48
  - 3.2.5 Integrated *FØ* tracking algorithm (IFTA) . . . 50
  - 3.2.6 Super resolution *FØ* determinator (SRFD) . . . 52
  - 3.2.7 FDA summary . . . 55
- 3.3 Syllabification . . . 55
- 3.4 The composite structure of *FØ* contours . . . 57

3.5 Automatic transcription of prosodic structure 60
3.6 Summary 68

4 Duration measures **71**
4.1 Database description 71
4.2 Phonetic units 73
4.3 Normalisation techniques 74
4.4 Evaluation of duration features 82
4.5 Optimisation of a relative duration feature 90
4.6 Summary 93

5 Energy measures **95**
5.1 Low-band energy contour 95
5.2 Energy features 98
5.3 Evaluation of energy features 102
5.4 Optimisation of a relative energy feature 108
5.5 Summary 112

6 Vowel quality measures **114**
6.1 Vowel context and speaker normalisation 115
6.2 Vowel target model 117
6.3 Vowel stability measure 125
6.4 Phonological rules 127
6.5 Summary 127

7 Fundamental frequency extraction **130**
7.1 Enhanced super resolution *FØ* determinator (*e*SRFD) 131
7.2 Evaluation of *FØ* determination algorithms 136
7.2.1 A laryngeal frequency tracker 137
7.2.2 Comparison of asynchronous frequency contours 140
7.2.3 Results and discussion 143
7.3 *FØ* contour post-processing 147
7.3.1 Non-linear smoother 147
7.3.2 De-step *FØ* filter 150
7.4 Summary 153

8 Automatic Prosodic Analysis **155**
8.1 Syllabification 157
8.1.1 Automatic syllabification from acoustic-phonetic parameters 157
8.1.2 Syllabification based on phonological rules 159
8.1.3 Evaluation procedure 160
8.2 Processing of *FØ* for an acoustic-phonetic description of intonation 164
8.2.1 Linear piece-wise stylisation of an *FØ* contour 164
8.2.2 Prosodic schematisation of a stylised *FØ* contour 167
8.3 Combination of acoustic parameters 172
8.4 Overview of the integrated prosodic analysis system 175
8.5 Performance evaluation of prosodic analysis algorithms 178

8.5.1 Integrated prosodic analysis (PCB-system) 179
8.5.2 Bayesian classifier (AW-system) 180
8.5.3 Knowledge-based rules approach (JLH-system) 182
8.5.4 Comparison 183
8.6 Conclusions 187

9 Application to Computer Aided Pronunciation Teaching **189**
9.1 Framework of the SPELL system 189
9.2 Example of automatic prosodic analysis applied to the SPELL framework 191
9.3 Further work 194

10 Summary and Conclusions **196**

A Least Median of Squares Regression **203**

B Bayesian Classification **206**

C Publications **210**

D MRPA and IPA Symbols **235**

References **238**

List of Figures

2.1 Syntactic and phonetic variations of *orange juice carton*. 10
2.2 Phonetic realisations of the Italian word *turbine** 13
2.3 Degrees of stress in *incompatibility*. 18
2.4 Finite-state grammar for **H/L** tone sequences 23
2.5 Prominent syllables and intonation features 28

3.1 Cepstrum-based *FØ* determinator (CFD) 40
3.2 Harmonic product spectrum (HPS) 43
3.3 Feature-based *FØ* tracker (FBFT) 46
3.4 Parallel processing method (PP) 49
3.5 Pitch accent decision filter 65

4.1 Syllable structure 74
4.2 Vowel durations 78
4.3 Consonant durations 79
4.4 Duration feature evaluation: Entropy score ($U_{1,\ldots,5}$) 84
4.5 Duration feature evaluation: Entropy score ($M_{0,\ldots,4}$) 85
4.6 Duration feature evaluation: Entropy score ($P_{0,1}$) 86
4.7 Duration feature evaluation: Entropy score ($S_{0,1}$) 87
4.8 Duration feature evaluation: Entropy score ($F_{0,1,2}$) 88
4.9 Optimisation of $K_{duration}$ threshold 92

5.1 Formation of a phone-level low-band energy contour* 97
5.2 Vowel low-band energies 100
5.3 Consonant low-band energies 101
5.4 Energy feature evaluation: Entropy score ($U_{1,\ldots,5}$) 103
5.5 Energy feature evaluation: Entropy score ($M_{0,\ldots,4}$) 104
5.6 Energy feature evaluation: Entropy score ($P_{0,1}$) 105
5.7 Energy feature evaluation: Entropy score ($F_{0,1,2}$) 106
5.8 Optimisation of K_{energy} threshold 109

6.1 Vowel targets in prominent and non-prominent syllables 120
6.2 Vowel quality feature evaluation 123

*The phonemic transcriptions use MRPA symbols — see Appendix D.

7.1 Analysis segments for the enhanced super resolution *FØ* determinator (*e*SRFD) 132
7.2 Example set of $p'_{x,y}(n)$ from previous frame 135
7.3 Glottal pulses evident in laryngograph data 138
7.4 Histograms of laryngeal frequency for male and female speech 139
7.5 Comparison of asynchronous *Fx* and *FØ* contours 142
7.6 FDA evaluation: Male speech 145
7.7 FDA evaluation: Female speech 146
7.8 Non-linear smoother 148
7.9 A de-step *FØ* filter and non-linear filter 151
7.10 Evaluation of post-processed *FØ* contours (for *e*SRFD) 154

8.1 Syllabification of the word *international** 162
8.2 Linear piece-wise stylisation of an *FØ* contour* 168
8.3 Transformation from a raw *FØ* contour to a schematic acoustic-phonetic representation of intonation* 171
8.4 Multi-way confusions across acoustic parameters 174
8.5 Block diagram of integrated prosodic analysis system 176
8.6 Parameter distributions for Hieronymus algorithm 184

9.1 Example of prosodic analysis in pronunciation teaching 193

List of Tables

4.1 Duration: Broad and fine phonetic classes 80
4.2 Duration: Confusion matrix for peak-picking algorithm 93
4.3 Duration: Confusion matrix for Bayesian classification 93

5.1 Low-band energy: Broad and fine phonetic classes 99
5.2 Energy: Confusion matrix for peak-picking algorithm 111
5.3 Energy: Confusion matrix for Bayesian classification 111

6.1 Vowel quality: Confusion matrix for peak-picking algorithm 122

8.1 Comparison of syllabifications based on phonological rules (by hand) and on acoustic-phonetic parameters (automatic) 163
8.2 *FØ* schema: Confusion matrix for pitch accent decision filter 173
8.3 Duration feature normalisation statistics 180
8.4 Energy feature normalisation statistics 181
8.5 PCB-system: Confusion matrix of prosodic transcription 182
8.6 AW-system: Confusion matrix of prosodic transcription 182
8.7 JLH-system: Confusion matrix of prosodic transcription 185
8.8 Cross-performance matrices of algorithm pairs 186

D.1 IPA/MRPA Symbols 237

Glossary of Technical Abbreviations

♀	female speaker.
♂	male speaker.
"–, \, /, ∨, ∧"	*FØ* trajectory descriptors.
Δf	*FØ* step threshold in pitch accent decision filter.
$\mu_E(unit\_type)$	mean phone-level low-band energy for unit-type.
$\mu_T(unit\_type)$	mean duration for unit-type.
$\sigma_E(unit\_type)$	standard deviation phone-level low-band energy for unit-type.
$\sigma_T(unit\_type)$	standard deviation duration for unit-type.
$\varphi_E^p(unit\_type)$	p'th-percentile phone-level low-band energy for unit-type.
$\varphi_T^p(unit\_type)$	p'th-percentile duration for unit-type.
$\varphi_Q^{(100-p)}(vowel\_type)$	$(100-p)$'th-percentile of the quadratic discriminant scores for vowel-type.
"*a*"	non-nuclear accented syllable (secondary stress).
ADC	analogue-to-digital converter.
"*Ap*"	syllable with maximum *FØ* in schema.
"*ap*"	accented syllable according to pitch accent decision filter.
$C_{1,2,3}$	features of formant frequency trajectories.
CFD	cepstral-based *FØ* determinator.
$D_{feature}$	duration feature corresponding to maximum classification rate.
DELTA	demonstrate, evaluate listening, teach and access (teaching paradigm).
$E_{feature}$	energy feature corresponding to maximum classification rate.
*e*SRFD	enhanced super resolution *FØ* determinator.
$F_{0,1,2}$	normalisations for the number of phones in a unit.
F_{Bark}	frequency value on a Bark-scale.
F_{Hz}	frequency value in Hertz.
F_{Mel}	frequency value on a Mel-scale.
F_s	sampling frequency.
FØ	fundamental frequency.
F1, *F2*, *F3*, *F4*	formant frequencies.
FBFT	feature-based *FØ* tracker.
FDA	*FØ* determination algorithm.
FFT	(fast) Fourier transform.

FIR	finite impulse response.
Fx	laryngeal frequency.
H	high tone.
H_0	null hypothesis.
$h(n)$	Hann window coefficient.
HMM	hidden Markov model.
HPS	harmonic product spectrum based method of *FØ* extraction.
i_{tp}	intercept of linear regression line associated with turning-point.
IFTA	integrated *FØ* tracking algorithm.
IPA	International Phonetics Association.
$K_{duration}$	duration threshold for "peak-picking" technique.
K_{energy}	energy threshold for "peak-picking" technique.
L	decimation factor.
L	low tone.
l_{Hann}	length of Hann window.
l_{median}	length of median filter.
LMedS	least median of squared residuals (linear regression).
LPC	linear prediction coding.
Lx	laryngograph signal.
$M_{0,\ldots,4}$	types of measuring duration and energy in a unit.
MRPA	machine readable phonemic alphabet.
"n"	nuclear accented syllable (primary stress).
n_0	fundamental period (in number of samples).
n_m	m'th fundamental period candidate (in number of samples).
N_{max}	maximum fundamental period (in number of samples).
N_{min}	minimum fundamental period (in number of samples).
'NA'	not available.
"NP"	non-prominent syllable ("u").
"P"	prominent syllable ("n"∪"a"∪"s").
p	percentage of non-prominent syllables in training data.
$P_{0,1}$	use of fine and broad phonetic classification.
$p_{x,y}(n)$	normalised crosscorrelation coefficient between sections x_n and y_n with decimation.
"PA"	pitch accented syllable ("n"∪"a").
PP	parallel processing method of *FØ* extraction.
Q_{vowel}	quadratic discriminant score for vowel.
Q'	vowel-type normalised quadratic discriminant score.
$q(n_m)$	normalised crosscorrelation coefficient between two sections of length n_M spaced n_m apart.
$r_{x,y}(n)$	normalised crosscorrelation coefficient between sections x_n and y_n.
"s"	unaccented but stressed syllable (tertiary stress).
$S_{0,1}$	measurements without and with the use of smoothing.
S_i	duration index.
s_N	set of data containing N samples.
s_{tp}	slope of linear regression line associated with turning-point.
SCRIBE	spoken corpus recordings in British English.

"*Sd*"	syllable with maximum duration of contour.
"*Se*"	syllable with maximum energy of contour.
"*sd*"	syllable corresponding to a local peak in the duration contour greater than $K_{duration}$.
"*se*"	syllable corresponding to a local peak in the energy contour greater than K_{energy}.
SPELL	interactive system for spoken European language training.
"*sq*"	syllable corresponding to a local peak in the vowel quality contour less than 0.0.
SRFD	super resolution *FØ* determinator.
$t_{interval}$	interval between successive analysis frames.
T_{srfd}	voicing classification threshold for SRFD.
$\overline{T_s}(i)$	average duration of prominent units which consist of i phones.
$\overline{T_u}(i)$	average duration of non-prominent units which consist of i phones.
T_{unit}	measured duration of unit.
ToBI	tone and break indices.
tp	a turning-point.
"*u*"	unstressed syllable.
$U_{1,\ldots,5}$	types of phonetic unit (nucleus, rhyme, lhyme, syllable, nucleus-to-nucleus unit).
"*UA*"	unaccented syllable ("*s*"∪"*u*").
"*ud*"	unstressed syllable on the basis of duration.
"*ue*"	unstressed syllable on the basis of energy.
"*up*"	unaccented syllable according to pitch accent decision filter.
"*uq*"	unstressed syllable on the basis of vowel quality.
$V_{1,\ldots,8}$	measures of the degree of spectral change in a vowel.
x_n, y_n, z_n	three consecutive sets of data, each containing n samples.
z_{mean}	z-score normalised value based on mean value.
$z_{percentile}$	z-score normalised value based on p'th-percentile value.

Chapter 1

Introduction

The aim of the work presented in this thesis is to automatically analyse the prosodic events in utterances of English spoken by native and non-native talkers. The automatic analysis concentrates on generating an acoustic-phonetic representation of sentential stress and intonation. An acoustic-phonetic representation of prosody has a practical application as a tool for computer aided pronunciation teaching.

Learning a foreign language involves acquiring an extensive knowledge of many different aspects of language. Prosody is one of these aspects of language. Prosody refers to the modulation of acoustic characteristics of speech above the level of phonemic segments in order to convey linguistic and paralinguistic information. This thesis is concerned with two prosodic features — stress and intonation. The term *stress* refers to the relative perceptual prominence of speech units larger than phonemic segments. The term *intonation* refers to the manipulation of pitch for linguistic and paralinguistic purposes above the level of phonemic segments.

In Chapter 2, it is argued that prosodic aspects of speech need to be explicitly taught to students who wish to communicate competently in a foreign language. A foreign language learner needs to acquire an understanding of native speech patterns and the ability to approximate near-native pronunciation because these influence both the student's comprehension and intelligibility. In particular, the language-dependent nature of stress and intonation is illustrated in Section 2.1 by examples of their functional differences and the way in which their acoustic-phonetic realisations differ in exemplar languages — English and Italian.

Computer aided pronunciation teaching involves automatic analysis of the speech of

a non-native talker in order to provide a diagnosis of a learner's pronunciation in comparison with the speech of a native talker. In order to offer a diagnosis of a learner's pronunciation with respect to prosodic features, it is necessary to describe the prosodic composition of the learner's speech and to determine any linguistically significant deviations from a near-native pronunciation.

A review of some phonological theories of prosody which aim to describe stress patterns and intonation in English is presented in Section 2.2. Current phonological theories of prosodic aspects of speech are language-specific. The prosody of the speech of a foreign learner of English, however, can be highly influenced by the prosody of a student's mother-tongue and by a student's stereotypic images of prosody in the English language, thereby rendering a language-specific phonological representation of prosody inappropriate in describing the prosodic aspects of the speech of a foreign language learner. Moreover, languages vary in the way acoustic parameters are modified to manifest prosodic aspects of speech. Thus, it is argued in Chapter 2, that the only secure means available to describe prosody in foreign language teaching lies in an acoustic-phonetic representation.

The automatic analysis of speech in computer aided pronunciation teaching uses a digitally sampled acoustic waveform as the only input parameter. An acoustic-phonetic representation of the prosodic aspects of speech is derived from this signal. Chapter 3 reviews a wide range of research relating to the identification, the extraction and the analysis of acoustic parameters which are correlated with prosodic features.

Research conducted to identify the acoustic parameters required for prosodic analysis is reviewed in Section 3.1. Measures of duration, energy, vowel quality and fundamental frequency ($F\emptyset$) are related to sentential stress and intonation. The problem of extracting the fundamental frequency from an acoustic speech signal is extensively addressed in the literature, relative to the problems of extracting the other three acoustic parameters. The functionalities of a selection of fundamental frequency determination algorithms are reviewed in detail in Section 3.2, demonstrating the complexity and diversity of the methods.

Syllables somewhat awkwardly fit into prosodic analysis. The definition of a syllable is associated with the relative degrees of the sonority of phonemic segments. Energy is an acoustic correlate of the sonority of phonemic segments. However, the energy

measures related to syllable prominence are dependent upon the definition of a syllable, thus creating a circular dependency. In addition to taking on a low-level role in the extraction of acoustic measures such as duration (for example, the measured duration of a syllable), syllables are used as the fundamental domain for prosodic descriptions. Thus, the automatic identification of syllables in connected speech forms an important part of prosodic analysis. A number of algorithms to partition speech into syllabic units are reviewed in Section 3.3.

In Chapter 3, it is asserted that the process of locating prominent syllables is complicated both by the interaction of acoustic features (duration, energy, vowel quality and fundamental frequency) in the manifestation of stress and by the fact that each of the four acoustic features is influenced by factors other than stress. Therefore, these acoustic features do not co-exist in a simple relationship to represent stress. In addition, the fundamental frequency of speech is affected by the talker's anatomy and physiology (macroprosody), the segmental content of an utterance (microprosody), cycle-to-cycle jitter (*FØ* perturbations) generated at the vocal folds, and erroneous *FØ* estimates generated by malfunctions in an *FØ* determination algorithm. Therefore, a raw *FØ* contour does not in itself form a complete representation of intonation. The composite structure of fundamental frequency is discussed in Section 3.4.

Moreover, there are a number of specific problems in measuring acoustic features that are required to automatically locate prominent syllables in connected speech. It is not evident from the reviewed literature how these acoustic features are best determined, over what phonetic domains the acoustic features are related with respect to prosody, or how they are best normalised for non-prosodic aspects of speech. A principal aim of this thesis is to address these problems.

Former algorithms proposed to automatically transcribe the prosodic structure of an utterance from acoustic parameters are reviewed in Section 3.5.

Chapters 4, 5, 6 & 7 concentrate on the derivation of duration, energy, and vowel quality features which can be used to automatically locate prominent syllables in connected speech, and on the extraction of fundamental frequency from an acoustic speech signal.

The domain of phonetic units whose duration and energy are to be determined as

optimal correlates of stress and normalisation techniques applied to them are investigated in Chapters 4 & 5. A number of measures of vowel quality are proposed in Chapter 6. Vowel quality measures are based on the assumptions that prominent syllables are well articulated and are less affected by contextual assimilation than non-prominent syllables, and that these properties are reflected in the nuclei of syllables. The underlying principle of these investigations is to normalise the acoustic parameters for non-prosodic aspects of speech such that the processed acoustic parameters can then be combined to form a prosodic description of speech. The data used in the experimental investigations is described in Section 4.1.

Methods to automatically determine the fundamental frequency of a speech waveform are addressed in Chapter 7. The most reliable and accurate method of determining the fundamental frequency is sought in order to minimise the number of errors occurring during $F\emptyset$ extraction from propagating into the prosodic analysis. An enhanced super resolution $F\emptyset$ determination algorithm (eSRFD) is proposed in Section 7.1.

Section 7.2 presents a comparative evaluation of the eSRFD algorithm and the $F\emptyset$ determination algorithms summarised in Section 3.2. The evaluation is performed by comparing $F\emptyset$ estimates with laryngeal frequency (Fx) estimates. An algorithm is presented which enables Fx contours to be generated from laryngograph data recorded simultaneously with speech. The modifications made to the super resolution $F\emptyset$ determination algorithm are shown to radically reduce the number of gross $F\emptyset$ doubling and halving errors and to improve the classification of voiced and unvoiced sections of speech. A novel de-step filter is proposed in Section 7.3.2 to post-process an $F\emptyset$ contour generated by an $F\emptyset$ determination algorithm in a way which further reduces the occurrence of errors. The task of reducing the occurrences of $F\emptyset$ perturbations and extraction errors is largely accomplished by the post-processing techniques described in Section 7.3.

The features of duration and energy, and the fundamental frequency extracted from the speech waveform need to be abstracted to form an acoustic-phonetic representation of sentential stress patterns and intonation in a syllabic domain. This is the aim of the automatic prosodic analysis system presented in Chapter 8.

The syllabification of connected speech forms a central part of automatic prosodic analysis, taking on three roles. Firstly, the prosodic aspects of speech are described in a

syllabic domain. Stress refers to the relative perceptual prominence of syllables and the description of intonation involves the association of pitch accents with prominent syllables. Secondly, the extraction of acoustic parameters is dependent upon the definition and identification of syllables. The duration feature found to optimally correlate with stress in Chapter 4 is dependent upon identifying the syllable lhymes[1] in an utterance. The optimal energy feature determined in Chapter 5 is dependent upon identifying the syllable nuclei in an utterance. The third role of the syllabic domain is related to the integration of the acoustic parameters. The syllabic domain inherently encompasses energy and segmental information which can be passed on to the analyses of energy, duration and *FØ* measures. For example, the location of syllable nuclei can serve as potential islands of reliability in the extraction of fundamental frequency. *FØ* estimates made within syllable nuclei are generally reliable because the speech signal within syllable nuclei is quasi-periodic and has a relatively high signal-to-noise ratio. These conditions are generally well suited to the assumptions made by *FØ* determination algorithms.

Section 8.1 compares the automatic syllabification of an utterance from its phonetic realisation, and the syllabification of an utterance on the basis of a set of abstract phonological rules. The proposed automatic syllabification algorithm is unique in its use of both a low-band energy contour and the segmentation (phone boundary and label information) of an utterance.

It is proposed in Chapter 8 that an acoustic-phonetic representation of intonation can be derived from a post-processed *FØ* contour by removing components of the contour which are due to microprosody, cycle-to-cycle jitter, and speaker dependent *FØ* range.

The stylisation of an *FØ* contour aims to prevent microprosodic variations from being confused as pitch accents, and hence effectively isolate the microprosodic and intonational components. The process of piece-wise linear stylisation of an *FØ* contour is presented in Section 8.2.1. Syllable information is used to aid the stylisation of an *FØ* contour since syllable nuclei locations are related to reliable sections of an *FØ* contour and microprosodic variations tend to dominate an *FØ* contour within the vicinity of short syllable nuclei in the context of unvoiced consonants.

[1]The *lhyme* of a syllable is defined in this thesis as the composition of syllable onset consonants and a syllable nucleus — see Section 4.2.

A stylised contour may contain some microprosodic variations given that the stylisation process will not be faultless. A stylised contour also contains macroprosodic and intonational components. The stylised contour must be processed with respect to the syllables of an utterance in order to eliminate any remaining microprosodic variations, to compensate for macroprosodic effects and to form an acoustic-phonetic representation of intonation. A process is proposed in Section 8.2.2 which aims to manipulate a stylised contour into a schematic form, with these goals in mind. A schematic representation of an *FØ* contour (an *FØ* schema) is independent of a speaker's fundamental frequency range and is derived so as to exhibit only the intonational component of the contour. The *FØ* trajectories of a schema can be associated with syllables to locate the syllables which are pitch accented. The integration of syllable information with the stylisation and schematisation of an *FØ* contour is an innovative approach.

The modular analyses proposed in this thesis are integrated to form a prosodic analysis system which generates an acoustic-phonetic description of the intonation and the sentential stress patterns of speech, in a syllabic domain. An overview of the system is described in Section 8.4 and its performance is evaluated in Section 8.5 relative to two algorithms formerly proposed in the literature.

In Chapter 9, the integrated prosodic analysis system is shown to produce prosodic descriptions which are useful in comparing the prosodic aspects of the speech of a non-native learner of English with the speech of a native English talker.

Chapter 2

Prosodic Description in Foreign Language Teaching

Learning a foreign language is a complex task which involves acquiring an extensive knowledge of the many different aspects of language, including vocabulary, grammar, and pronunciation. The relative weighting given to each aspect varies amongst teaching practices, but in all cases an understanding of native speech patterns and the ability to approximate near-native pronunciation are important components of foreign language learning, influencing both the student's comprehension and intelligibility (Kenworthy, 1987). Pronunciation of a foreign language is not only concerned with the articulation of phones (particularly those absent from the mother-tongue) and coarticulatory effects, but also with the modulation of acoustic features related to sequences of allophones; that is to say, with *suprasegmental* phenomena. The suprasegmental phenomena which form the subject matter of this thesis are the aspects of speech referred to as *prosody*. In particular, this thesis is concerned with two prosodic features — stress and intonation. The term *stress* refers to the relative perceptual prominence of speech units larger than phonemic segments. The term *intonation* refers to the manipulation of pitch for linguistic and paralinguistic purposes above the level of phonemic segments.

In this Chapter, it is argued that a need exists for the prosodic aspects of speech to be explicitly taught to students who wish to communicate intelligibly in a foreign language (Section 2.1). The argument will concentrate on two prosodic phenomena — sentential stress and intonation. The language-dependent nature of these prosodic phenomena will be illustrated by examples of their functional differences and the way in which their realisations differ in exemplar languages.

In order to teach the prosody of a language to a student, it is essential to understand and describe the suprasegmental phenomena exhibited in that language. It is also important to describe the prosodic composition of the student's speech, in order to determine *significant* deviations from a near-native pronunciation, and to offer some kind of corrective diagnosis. A review of some phonological theories of prosody which aim to describe stress patterns and intonation in English is presented in Section 2.2. Current phonological theories of prosodic aspects of speech are language-specific. The prosody of the speech of a non-native talker, however, can be highly influenced by the prosody of the source language[1] and by prosodic stereotypes of the target language. This renders a language-specific phonological representation of prosody inappropriate in describing the prosodic aspects of the speech of a foreign language learner. Thus this Chapter argues that the only secure means available to describe prosody in foreign language teaching lies in an acoustic-phonetic representation. There is, however, a problem in being able to determine which deviations in the acoustic-phonetic representation are *significant* at a linguistic level. The derivation of a non-language-specific relation between the phonetics and phonology of prosody is beyond the scope of this thesis which instead focuses on the need for an acoustic-phonetic description of prosodic aspects of speech.

2.1 The need for prosody in foreign language teaching

The goal of teaching any aspect of pronunciation in a foreign language course is not necessarily to render the student's pronunciation indistinguishable from that of a native speaker, but can instead be described as minimising the number of times when the student is misunderstood because a native listener erroneously interprets the intended meaning of the student's speech (Rivers, 1975; Harmer, 1983; Madsen, 1983; Kenworthy, 1987). This goal is coupled with teaching the mechanisms of pronunciation to students in order to help them to understand the speech of native speakers.

The following sections will demonstrate the importance of suprasegmental phenomena in language teaching for both a student's comprehension and intelligibility, high-

[1]The *source language* refers to the first language acquired by the student (ie. the language of the student's mother-tongue). The *target language* is the foreign language which is being taught to the student.

lighting how stress and intonation differ between languages in their functionality and realisation. This will support the assertion that prosodic aspects of speech need to be taught explicitly in foreign language courses.

2.1.1 Stress

The suprasegmental phenomenon, stress, refers to the relative perceptual prominence of a syllable or word in a particular context. Stress is subject to several different linguistic factors. It can be divided into a number of categories; for example, *lexical*, *grammatical*, *emphatic* and *contrastive* stress[2].

Lexical stress refers to the positioning of relatively prominent syllables in words spoken in isolation, as described by the lexicon of the language. In French, the primary lexical stress is placed invariably on the last pronounced syllable (excluding final *es*) with possible secondary (weaker) stresses occurring in polysyllabic words usually being positioned in alternate syllables prior to the primary stress (Martineau & McGrivney, 1973; Tranel, 1987). In English, however, the position of lexical stress (primary or secondary) is dependent upon the word chosen. The principles which govern the placement of lexical stress in English words have previously been investigated in depth (Kingdom, 1958; Chomsky & Halle, 1968; Fudge, 1984). For the purposes of optimal intelligibility of a student's speech, it is important for lexical stress to be placed on the correct syllables of a word, as pronounced by a native speaker.

Grammatical stress refers to the lexical stress patterns which appear in phrases of connected speech, as governed by the grammar of the language. In English, the content words (though not necessarily all of them) maintain their lexical stress patterns in connected speech, whereas the function words do not (although they may carry contrastive stress). In French, grammatical stress is only associated with the last pronounced syllable in a rhythmic group (Price, 1991). Syntactic distinctions can be made by distributing grammatical stress at different strengths through an utterance. For example, a distinction can be made between "**O***range juice* **car***ton*" (with the strongest grammatical stress being applied to the first syllable of *orange* and the weakest stress to the

[2]This list of categories is not intended to be complete, as the different possible types of stress are dependent upon the language spoken. The illustration here is concerned only with some stress types found in French, Italian, and English.

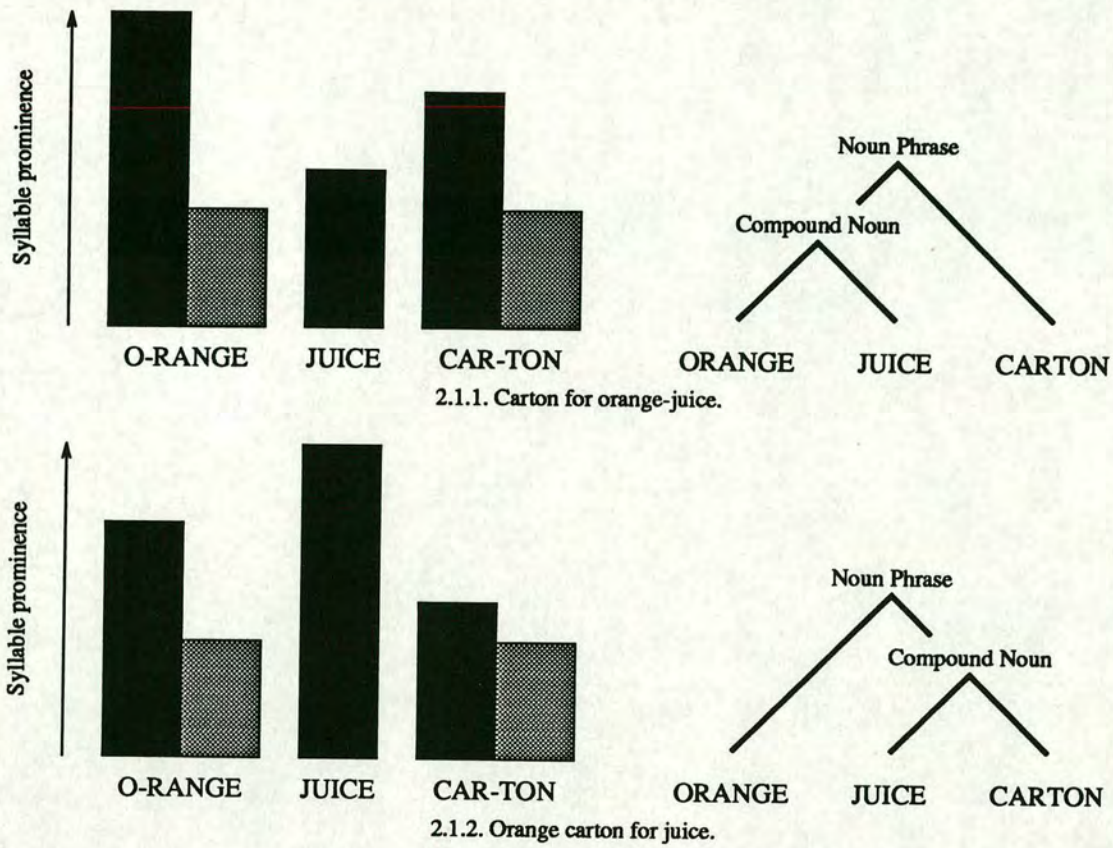


Figure 2.1: Syntactic and phonetic variations of *orange juice carton*.

word *juice*) meaning, 'a carton for orange-juice'; and "*orange* **juice** *carton*" (with the strongest grammatical stress on *juice* and the weakest stress on *carton*) meaning, 'an orange coloured carton which is used to contain juice'. These two syntactic variations of the phrase *orange juice carton* and their corresponding patterns of syllable prominence are illustrated in Figure 2.1. Accounts of the principles behind the relative strengths of prominent syllables in such compounds are given by Chomsky & Halle (1968) and Fudge (1984). Improper use of the relative strengths of grammatical stress by a non-native speaker may result in syntactic ambiguities.

A syllable may also be made prominent in order to place emphasis on a particular word. This type of stress is referred to as emphatic stress. In English, such stress is realised by giving additional prominence to syllables already carrying grammatical

stress; whereas in French, this effect tends to be produced by applying prominence to the first syllable of a word beginning with a consonant, or by the articulation of a glottal closure before a word beginning with a vowel (Martineau & McGrivney, 1973; Price, 1991). This may pose difficulties for English speakers of French. For instance, an English speaker may pronounce "**mer***veilleux*" with stress on the first syllable as in its English equivalent "**mar***vellous*" (example from MacCarthy, 1975). This type of mother-tongue interference may be interpreted as emphatic stress by a French listener, although emphasis is not intended by the speaker.

Contrastive stress refers to the positioning of prominent syllables for semantic clarification. In English, contrastive stress can be placed on syllables where grammatical stress does or does not already occur, and in both function and content words. It may be used to form an explicit contrast between phonetically related words; for example, "*I said int***ra***ctable, not int***er***actable*;" or to make an implicit contrast; for example, "*He was going* **to** *the shop*," rather than, perhaps, "...**from** *the shop*".

The three types of stress which occur in connected speech, grammatical, emphatic, and contrastive stress, are collectively referred to in this thesis as *sentential* stress. The patterns in time of the syllables and sentential stress give rise to the *rhythm* of an utterance in some languages, such as English and Italian. The acquisition of rhythm is an important aspect of learning a foreign language, since incorrect rhythm can hinder a listener's ability to understand the segmental content of a student's speech (Cutler & Norris, 1988).

It can be seen from the above descriptions that the phonetic realisation of different types of stress is language dependent and that stress is important from both a syntactic and a semantic viewpoint. Therefore, in optimising pronunciation, foreign speakers must try to prevent reproducing the stress patterns of their mother-tongue on the target language (except where there is a common overlap in the stress patterns of the source and target languages). They must also understand the functionality of the different types of stress and learn how to realise them in the same way as a native speaker of the target language.

The perceived prominence of syllables is manifested by variations in duration, intensity, pitch and vowel quality (these features will be discussed in detail during the

following Chapters). Dauer (1983) proposes that languages vary in the way in which these acoustic characteristics of speech are modified in manifesting stress. For example, unstressed vowels are often reduced to schwa /ə/ or centralised /ɨ/ in English, whereas the quality of all vowels tends to be maintained in French and Italian. Furthermore, grammatical stress in French is much weaker than that of English and Italian, and is often overshadowed by emphatic or contrastive stress (Tranel, 1987; Price, 1991). Learning how to apply stress in a foreign language is therefore a task which is more complicated than just learning stress placement.

In Italian, lexical stress usually lies on the penultimate syllable of a word. If the stress of a word lies on the final syllable, it is marked as such in the lexicon. In some cases, the stress can lie on the anti-penultimate syllable. However, anti-penultimate lexical stress is not marked in the lexicon (Chapallaz, 1979). A student of Italian must therefore learn the lexical stress placement of Italian words, as with words in English. There are some homographic words in Italian where a shift in stress from the penultimate syllable to the anti-penultimate syllable is associated with a change in meaning — for example; *ancora* /'a ŋ k o r a/ 'anchor', /a ŋ 'k o r a/ 'again'; *compito* /'k o m p i t o/ 'task', /k o m 'p i t o/ 'poised'; *turbine* /'t u r b i n e/ 'whirlwind', /t u r 'b i n e/ 'turbines'[3].

Figure 2.2 shows the phonemically transcribed speech waveforms, low-band energy contours (see Section 5.1) and fundamental frequency (*FØ*) contours for the Italian word *turbine* spoken in its two lexical forms by a native Italian speaker, and spoken in an ambiguous form by a native speaker of English. The prominent syllables are pronounced by the Italian speaker with a pitch glide over the vocalic nuclei of the syllables and with a long syllable nucleus duration relative to the unstressed syllables. In the English speaker's attempt to pronounce the Italian word /t u r 'b i n e/ 'turbines', the ratio of the duration of the initial syllable *tur-* to the duration of the syllable *-bi-*, and the ratio of the peak low-band energy of the syllable *tur-* to the peak low-band energy of the syllable *-bi-* are greater than those of its Italian counterpart. This is analogous to the pronunciation of the English word *turbines* /'t ɜ b aɪ n z/, with the lexical stress on the initial syllable, and tends towards the pronunciation of the Italian word /'t u r b i n e/ 'whirlwind'. In

[3]Many thanks to Fabrizio Carraro, a native speaker of Italian, for providing these examples, and to Dr. Bob Ladd for pointing out that ambiguities can arise from the phonetic realisation of these words when spoken by a non-native speaker of Italian.

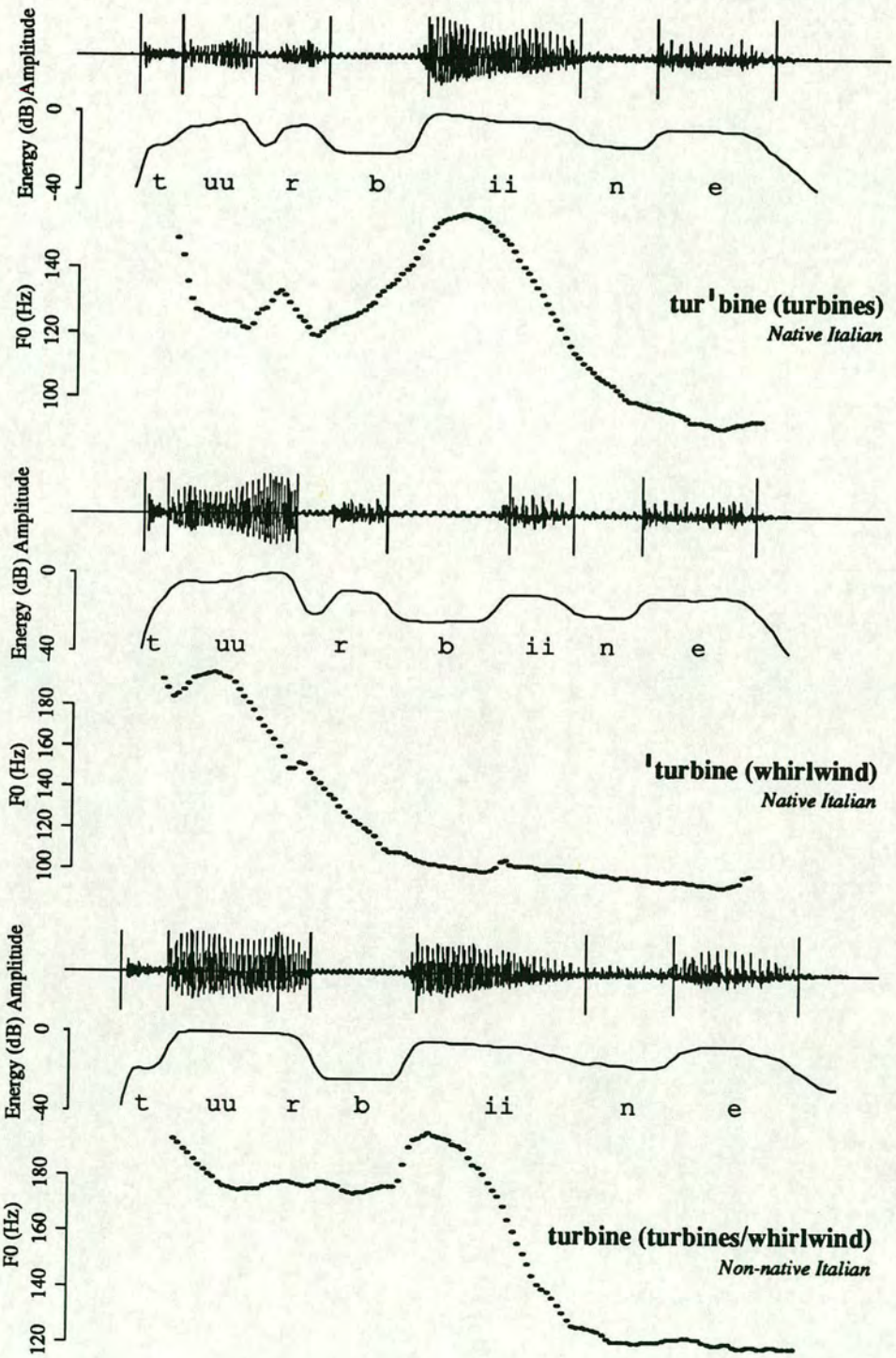


Figure 2.2: Phonetic realisations of the Italian word *turbine**.

English, however, a pitch glide over the vocalic nucleus of a syllable can act as a strong cue to syllable prominence and may outweigh any duration or energy cues (Fry, 1958). Thus, in the knowledge that the lexical stress of the Italian word /t u r 'b i n e/ 'turbines' is to be placed on the penultimate syllable, the English speaker produces a pitch glide over the syllable *-bi-*. As far as the non-native speaker is concerned, the lexical stress has been realised on the correct syllable, but the resultant utterance lies close to an Italian listener's limen between hearing /t u r 'b i n e/ 'turbines' and /'t u r b i n e/ 'whirlwind'. Thus, although a foreign speaker's knowledge of the phonological stress placement may be correct, the incorrect phonetic realisation of the stress can result in ambiguity.

2.1.2 Intonation

The second suprasegmental phenomenon considered in this thesis, intonation, refers to the variations in pitch which give an utterance a characteristic melody. Intonation is used to convey linguistic and paralinguistic information (Lehiste, 1970). However, many foreign language courses do not include the explicit teaching of intonation as an integral part of the syllabus.

> "The majority of course books which include overt instruction on intonation use the incidental, not the systematic approach. This means that intonation patterns are randomly selected, and do not exemplify intonational categories drawn from a linguistic description of intonation choices and their meanings. Typically, an intonation contour is simply presented for imitation, without any gloss as to meaning, or any contrast with other possibilities."
>
> (Brazil *et al.*, 1980, pp.115)

Brazil *et al.* (1980) suggest that the omission of formal intonation teaching may be attributed to the fact that students are expected to absorb intonational phonology implicitly through exposure to the speech of native speakers. Furthermore, many teachers take the view that emphasis should be placed on teaching grammatical aspects of a language and in expanding a student's active vocabulary, rather than on pronunciation training which may always be coloured by mother-tongue interference. It is assumed that, if a student does make a mistake in pronunciation, native listeners are usually capable of adapting to the foreign accent and can assert the meaning of a mispronounced utterance through contextual and linguistic cues. This, however, may not always be

true. For example, if a non-native English speaker asks a pharmacist, "Could I please have some medicine for my /k aʊ/ (*cow*)?" it is assumed that the foreigner will not be directed to a veterinary clinic, rather a native listener is expected to identify the mistake of pronouncing the word *cough* /k ɒ f/ in a similar manner to the word *plough* /p l aʊ/. The argument for omitting explicit teaching of pronunciation on the basis that native listeners can detect such graphemic/phonetic confusions is therefore flawed. Moreover the omission of explicit pronunciation teaching assumes that students only need to *be understood.* If students are to gain any communicative competence and confidence, they will also need *to understand* contrastive sounds made by a native speaker.

Although the previous example was concerned with a graphemic/phonetic confusion, misinterpretations can also arise through the use of different intonation patterns. For example, saying, "That's the bus to Edinburgh," with an interrogative intonation could result in a sarcastic reply, "Oh, really. I always wanted to know that," if a foreign listener misinterprets it as a declaration.

Prosody also plays a role in removing the ambiguity from phonetically similar but syntactically different utterances (Price *et al.*, 1991). For example, the sentence, "He saw that petrol can explode," may be interpreted either as meaning, 'he saw the explosion of that specific container of petrol'; or as meaning, 'he understood that petrol as a substance is capable of exploding'. In the former interpretation, a prosodic boundary would be placed immediately after *can*; whereas in the latter interpretation, the boundary would be placed before *can*. This shift in the placement of the prosodic boundary is realised by variations in pitch and duration. In both cases there is a close relation between the prosodic structure and the desired syntactic structure.

In order to understand a foreign language, students must have at least some knowledge of the contrastive intonational patterns in the language, even if they have difficulty in producing the sounds themselves. Furthermore, the learner of a foreign language will inherently have less experience of the possible grammatical, intonational and contextual variations of utterances than a native listener. The non-native listener will need as many cues as possible to interpret an utterance. A knowledge of intonational phonology would therefore be of assistance.

2.2 Phonological theories of prosody

Phonology is traditionally defined as the study of sound units in a language together with the study of the structure of the relationship between these units of sound. A student of a foreign language is, implicitly, also studying the sound patterns of a language. It should therefore be advantageous to make use of phonological theories in teaching a language. In particular, if the intonation and stress patterns of a specific language are to be taught to a student then it will be necessary to know how to describe intonation and stress patterns in that language. It will also be necessary to describe the intonation and stress patterns of the student's speech in order to determine *significant* deviations from a native-like prosody, and then to offer some kind of corrective diagnosis.

Speech is extremely complex and a very large number of different pitch contours and sequences of prominent syllables of different strength can exist. Phonology aims to reduce the complexity of speech by identifying patterns of sounds in a language.

Over the last half century (as early as Pike, 1945) researchers of phonology have addressed the relationships between prosodic events (intonation, stress, rhythm, and pausing patterns) and investigated their interaction with syntactic structure. Chomsky & Halle (1968) propose a variety of phonological rules to transform the syntactic representation of a sentence (*surface structure*) into a phonetic representation. However, noting that syntactic phrases do not always correspond with perceived phrasing in speech, they also introduce a number of 'readjustment' rules to relate the surface structure to 'phonological phrases' which differ from the syntactic phrasing. Liberman & Prince (1977) propose that a hierarchy of phonological structures, separate from syntactic structures, is required to adequately describe prominence relations amongst the words and syllables of a sentence. They argue that the differences in syntactic and phonological structures are confined to the level of the prosodic word and lower level constituents (the foot and syllable) and that syntactic and phonological structures are isomorphic above the level of the word. This view is challenged by Selkirk (1984), who presents a phonological hierarchy containing an intonational phrase, phonological phrase, prosodic word, foot and syllable, which is separate from but related to the syntactic structure.

Given the view that phonological and syntactic structures are separate from but related to each other, and that the relations between these two structures are not yet

fully understood, it is inappropriate to teach the prosody of a language in terms of how prosodic events are related to grammar. Intonation and stress need to be described in their own right. Their relation to syntax and semantics can be taken as a separate issue.

The following Sections will review some descriptions of stress and intonation in English. It will be argued that although a phonological description of English may or may not be adequate to describe the prosody of a native speaker, the language-specific nature of phonology theory prevents it from being used to describe the prosody of English spoken by a non-native speaker. The work considered here focuses on the need for a phonetic description of prosodic aspects of speech to be used in foreign language teaching.

2.2.1 Degrees of stress

In representing degrees of stress, Fudge (1984) follows the convention of Crystal (1969) which uses four degrees of stress distinguishable at an acoustic level.

> "The physical properties which signal stress in English do not enable hearers, even trained phoneticians, to distinguish consistently more than three degrees of strength [*plus an unstressed category*]." (my italics)
>
> (Fudge, 1984, pp.137)

The syllable which bears the main sentential stress in a phrase or sentence is referred to as the primary (or nuclear) stress (marked by the prefix '). The traditional view is that nuclear stress falls on the final pitch accented syllable in the phrase. Prominent (and pitch accented) syllables occurring before the nuclear stress are described as secondary stress (marked by the prefix ˌ). The third level of stress (marked by the prefix ˌ) refers to syllables which are usually pronounced with a full vowel rather than a reduced vowel, but which are not associated with a pitch accent as in the cases of primary and secondary stress. All other syllables are unstressed (and are not marked in any way). The occurrences of nuclear, secondary and tertiary stress in phrases are considered further in Section 2.2.2. These levels of stress can also be discerned at the word level. For example, in the word *incompatibility* (illustrated in Figure 2.3) the primary (or nuclear) lexical stress falls on the syllable *-bi-*, but weaker, secondary and tertiary stress fall on the syllables *in-* and *-pa-*, respectively. The remaining syllables may be pronounced with a reduced vowel (typically either a schwa /ə/ or a centralised /ɨ/) and are unstressed.

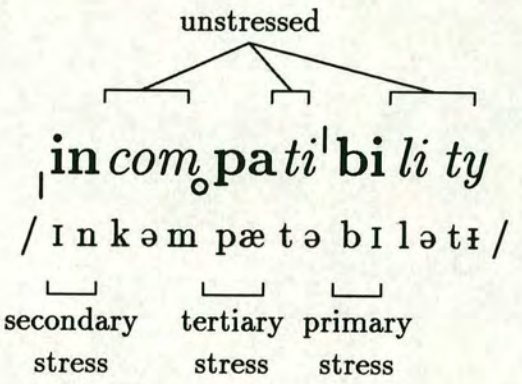


Figure 2.3: Degrees of stress in *incompatibility*.

An example was given in Section 2.1.1 of the syntactic differences which can be conveyed by changing the relative strengths of syllables in a phrase. Chomsky & Halle (1968) propose that the relative strengths of each syllable in a phrase may be predicted by applying a set of phonological rules to the syntactic structure of the phrase. Although, by inference, application of Chomsky's phonological rules may predict an unlimited number of degrees of stress, especially in phrases which have a complex syntactic structure, it is not suggested that each degree of stress will have an equivalent acoustic realisation. Chomsky's theory requires that a listener should be able to perceive the predicted stress pattern, since the predicted stress patterns describe a perceptual reality, whereas in fact a listener will only be capable of transcribing the stress pattern of an utterance when given a knowledge of the syntactic structure of the language to which it applies, since not all degrees of stress will be discernible from the acoustic speech signal. Lieberman (1965) shows that listeners capable of describing the stress pattern of a synthetic stimulus conveying only the acoustic correlates of stress (fundamental frequency, duration and energy) may represent the same acoustic properties quite differently when such acoustic dimensions are associated with the words of an utterance. He concludes that the listener may be inferring the presence of secondary and tertiary stress from knowledge of the grammatical attributes of the words of the utterance, rather than from acoustic

parameters. It is worth noting, however, that the synthetic stimuli used by Lieberman contain the same vowel in all syllables and the absence of reduced vowels in the synthetic stimuli may also be contributing to a listener's inconsistency in identifying secondary and tertiary stress.

It is considered that a learner of a foreign language is unlikely to have the same degree of knowledge of complex syntactic structures for that language as a native speaker. A foreign learner cannot, therefore, be expected to predict the stress pattern of a new phrase, even if equipped with a set of relatively simple phonological rules such as those proposed by Chomsky & Halle (1968). In teaching English stress, the only way to test if a student can produce an utterance with an intelligible stress pattern, is to listen to and analyse the acoustics of the student's speech waveform.

Thus, although there may be many linguistically significant gradations of stress in English, there are far fewer which are realised acoustically. A student of the English language should be able to produce stress patterns (manifested in the speech waveform) in a similar manner to a native speaker. For this reason, the descriptions here will be restricted to the three degrees of stress (plus an unstressed category) which are acoustically distinct. If a student is producing the correct stress pattern at an acoustic-phonetic level, then it can be assumed that the phonological rules which are being applied are adequate. It is not, therefore, necessary to deduce the phonological rules which a student uses, or to determine if these rules are the same as the rules employed by a native speaker.

2.2.2 Configuration theory

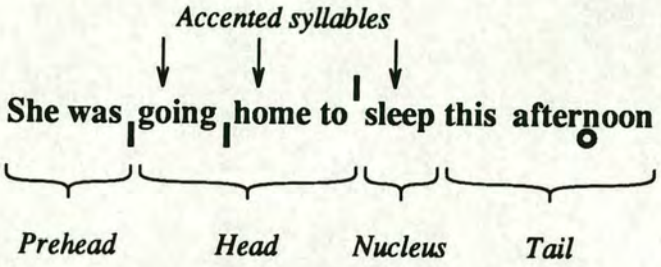
The intonation of an utterance can be described in terms of *tone units* (Crystal, 1969). The tone unit is regarded as the primary unit of intonational structure. Most theories of intonational phonology include some kind of primary unit of intonational structure which has also been called the *tone group* (Halliday, 1970), the *word group* (O'Connor & Arnold, 1973), and the *intonation phrase* (Pierrehumbert, 1980), although some differences exist in the detailed definitions of these constructs.

Crystal (1969) considers each tone unit to be composed of, at most, four components — the *prehead*, *head*, *nucleus* and *tail*. The nucleus is the only obligatory part of a tone

unit.

Tone unit = (*Prehead*) (*Head*) **Nucleus** (*Tail*)

For example,



The final, accented, prominent syllable of the tone unit forms the *nucleus* of the tone unit, and is associated with the *primary* stress (peak of prominence) in the tone unit. The *tail* consists of any additional syllables which may occur after the nucleus. The continuation and completion of the nuclear pitch movement carries into the tail. For this reason, Halliday (1970) treats the nucleus and tail as one unit, which he refers to as the *tonic*. The first syllable of the tonic is called the *tonic syllable* and is, by definition, prominent. In Halliday's description, the nuclear pitch movement associated with the prominence of the tonic syllable includes either a continuous rising and/or falling pitch *glide* on the syllable, or a pitch *jump* up or down occurring immediately before the syllable.

Crystal proposes four primary types of nuclear pitch movement, 'rise', 'fall', 'rise-fall', and 'fall-rise'. Each movement is phonetically represented at a number of different relative pitch levels (high onset, mid onset, low onset ...) and pitch ranges (wide, narrow). He also postulates the existence of a 'level' tone. There are also four permissible combinations of these primary tones which can occur within a single tone unit and which are claimed to be phonologically distinct. The compound nuclear tones proposed are 'rise' or 'fall-rise' plus 'fall', and 'fall' or 'rise-fall' plus 'rise'.

O'Connor & Arnold (1973) describe only seven categories of tone — 'high-rise', 'low-rise', 'high-fall', 'low-fall', 'rise-fall', 'fall-rise' and 'mid-level' — and discusses only one possible compound tone — 'fall-plus-rise'.

Halliday (1970) classifies pitch movements into five primary tone categories plus two primary compound tones which are made up of combinations from these five — tone **1** 'fall', tone **2** 'rise' or 'fall-rise (sharp)' involving a sudden change in direction from fall to

rise, tone **3** 'low rise', tone **4** 'fall-rise (rounded)' involving a smooth, gradual transition from fall to rise, and tone **5** 'rise-fall (rounded)'. The compound tones proposed by Halliday are tone **13** 'falling plus low rising' and tone **53** 'rising-falling (rounded) plus low rising'. There is noticeably no 'level' tone in Halliday's description.

The *head* (or *pretonic* in Halliday's terms) consists of the syllables from the first accented, prominent syllable in the tone unit up to, but not including, the tonic syllable. It is assumed in the above configuration theories that individual accented, prominent syllables in the head (*secondary* stress) do not carry any linguistic significance in the way that the tonic syllable does. The syllables bearing secondary stress are therefore treated and described as a single unit. O'Connor & Arnold distinguish four characteristic tone patterns in the head, each having a limited co-occurrence with the nuclear tones — a 'low' head which occurs only before a 'low-rise' nuclear tone, a 'high' head which occurs before all nuclear tones except the 'fall-rise' tone, a 'falling' head which occurs only before a 'fall-rise' nuclear tone, and a 'rising' head which occurs only before a 'high-fall' nuclear tone. Crystal and Halliday offer more comprehensive descriptions of the head. Crystal proposes four types of 'falling' head, two categories of 'rising' head, a 'falling-rising(-falling)' head and a 'rising-falling(-rising)' head. Unlike the description by O'Connor & Arnold, the co-occurrences of head and nuclear tones are unconstrained within Crystal's description.

The *prehead* represents any syllables bearing *tertiary* stress and unstressed syllables that exist prior to the first accented, prominent syllable (of the head or, if no head exists, of the nucleus) in the tone unit. O'Connor & Arnold describe the prehead as being either 'high' level or 'low' level. Crystal, however, proposes that the prehead can take four distinguishable pitch heights, and mentions evidence for the possibility of a fifth level.

There are a number of problems which arise from these theories of intonational phonology.

The three descriptions cited above are based on the intonation of speech from native speakers of (Southern British) English, however the intonation of non-native speakers of English will, in most cases, be strongly affected by mother-tongue interference. This means that the pitch configurations identified in the above descriptions of English into-

nation are probably inadequate to describe a foreigner speaker's intonation.

Although the nuclear pitch movement of a tone unit is presented by these three descriptions of intonation in terms of rising, falling or level pitch, the phonetic differences that are treated as linguistically important varies between them. This is particularly noticeable of the permissible compound tones. It is disturbing to find such disagreements over the phonological categorisation of pitch movements.

These descriptions of intonation can also be criticised for their inability to capture a hierarchical prosodic structure. Ladd (1986) argues that boundaries between prosodic constituents are loosely defined, yet they are intended to be made up of well defined internal phonological structures. In particular, Ladd argues that the definitions of tone unit and nucleus are somewhat interdependent. This makes it difficult to partition speech into a linear sequence of one type of prosodic entity and leads to inconsistencies between intonational transcriptions of a given utterance (Grice & Barry, 1991). This is also reflected in the inconsistent classification of compound tones by Crystal, O'Connor & Arnold, and Halliday.

> "The compound tones ... are really sequences of two tones which have, however, become fused into a single tone group, so that there is no possibility of introducing a pretonic between the two."
>
> (Halliday, 1970, pp.12)

It may be possible to draw a distinction between two "fused" tone groups by postulating a prosodic entity between the tone group and the prosodic word. This debate of prosodic structure will be revisited in the following Section.

2.2.3 Tone sequence theory

A phonological model of intonation proposed by Pierrehumbert (1980) describes a pitch contour as a series of interpolations between successive pitch targets. This model stems from earlier work related to the concept of pitch targets; in particular, that of Liberman (1975) and Bruce (1977). The pitch targets are the phonetic realisation of two phonologically distinct tone segments, a high tone (**H**) and a low tone (**L**). A distinction is made between pitch accent related tones, phrase accent tones and boundary tones.

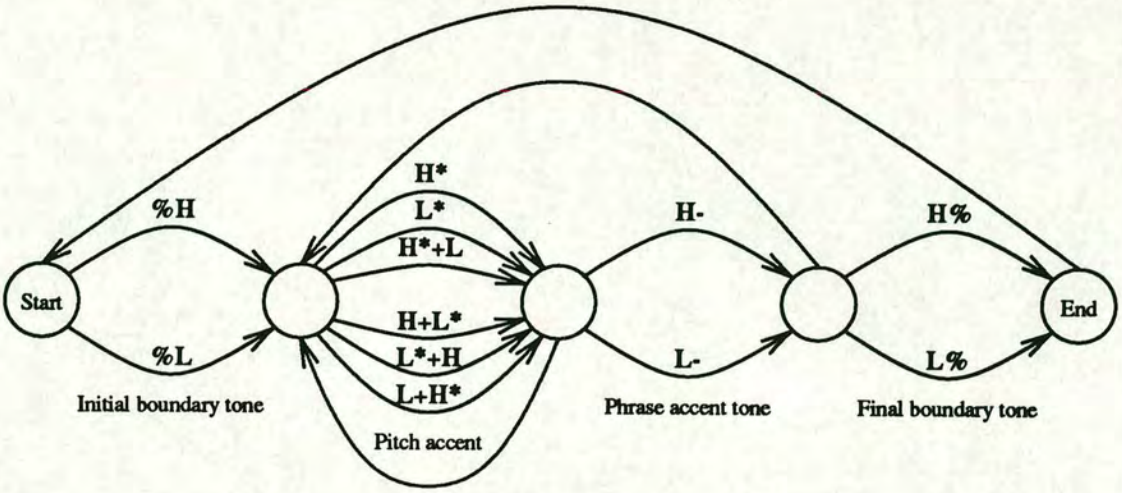


Figure 2.4: Finite-state grammar for **H/L** tone sequences (Pierrehumbert, 1980)

The permissible sequences of tone segments are specified by a finite-state grammar, as shown in Figure 2.4.

Strong evidence supporting the existence of pitch targets in preference to pitch movements (as in the configuration theory) is provided by Bruce (1977). Bruce finds a reliable correlate of Swedish word accent to be a local peak of fundamental frequency which is time-aligned to the accented syllable. In some circumstances, however, the rise up to the peak, and/or the fall after it, can be reduced, making it difficult to relate the accented syllable with a specific pitch movement. Bruce argues that reaching a certain pitch level at a particular time is an important correlate of word accent, not the pitch movement. A similar observation is also found in Crystal's (1969) argument for the existence of a 'level' tone, which is inconsistent with having defined nuclear tone as the most prominent pitch *movement* in a tone unit,

> "...in English there is often ...clear evidence of a tone-unit boundary, but no audibly kinetic tone preceding."
>
> (Crystal, 1969, pp.215)

Pitch accent related tones are designated to the location of syllables bearing sentential stress. In Pierrehumbert's analysis, pitch accents in English are described using single

and double tones. A *star* (*) diacritic is used to identify the tones associated with accented, prominent syllables. The six pitch accents are **H***, **L***, and the four bitonal accents, **H*+L**, **H+L***, **L*+H**, and **L+H***. Each pitch accent is realised across any number of adjacent syllables, referred to as a *prosodic word*. There are no restrictions on the type of pitch accent which can be associated with an accented, prominent syllable, as in Crystal's model (where accented syllables in the head are treated in a different way to those in the nucleus).

Phrase accent tones are specified by the grammar to exist only between the final tone related to a pitch accent and a boundary tone. They are marked by a *hyphen* diacritic (**H–**, **L–**) and denominate *intermediate phrases* (Beckman & Pierrehumbert, 1986).

Boundary tones are associated with a higher level of phrasing than the phrase accent tones. They denominate *intonational phrases* and are marked by a *percent* diacritic (**H%**, **L%**).

Empirical observations of the phonetic realisation of the high and low tones show the pitch targets to be of different height. It is believed that there are many factors which contribute to this effect. One common observation is an overall downward trend in pitch (or more specifically, in fundamental frequency) over the course of an intonational phrase. Pierrehumbert (1980) proposes that this global downdrift is mainly attributed to the phonological effect of *downstep* — a stepwise lowering of high pitch targets at moments controlled by the speaker. Long term downdrift (possibly over a number of intonational phrases, but not spanning an intake of breath) may also be attributed to a physiological property referred to as *declination* (Vaissière, 1983). Downstep is a reduction of pitch range that lowers the *FØ* realisation of any high tones subsequent to a *downstep trigger*. In Pierrehumbert's description of intonation, downstep is triggered by bitonal accents. The **H–** intermediate phrase accent triggers *upstep*. Downstep is marked explicitly by an *exclamation mark* diacritic (**!H**).

The notion of a "strict", layered, hierarchical structure of intonation (*strict layer hypothesis* (Selkirk, 1984)) is implicit within this theory. An intonational phrase consists of a number of intermediate phrases (and of no other type of prosodic constituent) which are in turn made up of pitch accents across a number of syllables (prosodic words). Beckman & Pierrehumbert (1986) suggest that there are at least two, and possibly

three[4], levels of prosodic phrasing above the prosodic word. In order to gain some insight into the depth of prosodic structure, the lengthening of phrase final syllables have previously been studied. Wightman *et al.* (1992) find that the duration of the vowel preceding a prosodic boundary correlates with the strength of the boundary, and that vowel duration variations at these boundaries indicate three levels of prosodic phrasing above the prosodic word. Ladd & Campbell (1991) show that the distribution of syllable durations can be more accurately modelled from data annotated by four levels of phrasing than by two-level intonational/intermediate phrasing. Ladd (1992) argues that such studies on the realisation of prosodic boundaries provide evidence that prosodic structure is of variable depth and, assuming that the variability in the phonetic realisation is not due to paralinguistic effects, argues that this is incompatible with the notion of a fixed (but unknown) depth of prosodic structure proposed in the strict layer hypothesis. As a possible solution to this problem, Ladd suggests that prosodic constituents of the same category can be nested. Thus, for example, an intonational phrase can be made up of other intonational phrases, thus forming a compound intonational phrase. Ladd's view of prosodic structure yields fewer constraints on the mapping of the syntactic domain to the prosodic domain than the strict layer hypothesis, and is more compatible with the notion of compound syntactic structure.

The need for some widely acceptable scheme of labelling prosodic aspects of speech has only recently been addressed. The conventional labelling of prosodic structure is language-specific and dependent on the prosodic theory employed (Barry & Fourcin, 1992). The task of labelling speech is split between giving the manifestations of speech a symbolic transcription and annotating the physical speech signal.

A transcription scheme for labelling prosodic aspects of speech in American English is proposed by Silverman *et al.* (1992) based on tones and break indices (ToBI). The perceived connectivity between adjacent words is represented by a five-value *break index*[5] (Price *et al.*, 1990; Price *et al.*, 1991). A lower value of break index corresponds to a greater degree of prosodic coupling between neighbouring words. The break indices

[4]Beckman & Pierrehumbert (1986) suggest the possibility of an *accentual phrase* level between the intermediate phrase and prosodic word levels.

[5]Price *et al.* (1990; 1991) propose a seven-level break index. Level-5 (boundary marking a grouping of intonational phrases) and level-6 (sentence boundary) break indices are submerged into level-4 break indices in the ToBI transcription scheme.

mark the boundaries within a clitic group (level-0), normal word boundaries (level-1), boundaries between minor groupings of words (level-2), intermediate phrase boundaries (level-3) and intonational phrase boundaries (level-4). The break indices are tied to the strict layer hypothesis.

The ToBI transcription scheme lies between being a phonological description and a phonetic description of speech. The three sequences of tone segments represented as **H* L–L%**, **L+H* L–L%** and **L*+H L–L%** each describe a rise-fall contour. The peak in *FØ* associated with the high tone occurs at different times relative to the prominent syllable in the three tone sequences. It is not clear whether these three tone sequences describe three phonological categories which are linguistically contrastive, or whether they describe three phonetic variations of a rise-fall contour.

Empirical observations of the **H+L*** bitonal pitch accent show the phonetic realisation of the low tone component to be higher than the low tone in the **L*** pitch accent and in the **H*+L**, **L*+H**, and **L+H*** bitonal accents. The **H+L*** accent is described as **H+!H*** in the ToBI transcription scheme to reflect this phonetic distinction.

The relative prominences of syllables within each word of an utterance are not marked in the ToBI transcription scheme. This categorical aspect of prosody is omitted because it is assumed that the relative prominence of syllables is predictable (for a native speaker of English) and that this stress information can be derived from a word's lexical entry. If the relative prominence of syllables is not as predicted then a native listener will judge the word as being mispronounced. There is, however, no way of transcribing this in the ToBI scheme.

The analysis of intonation in terms of high and low tone sequences has been applied to several languages other than (American) English; for example, Dutch (Gussenhoven, 1984), French (Merten, 1987), Hungarian and Romanian (Ladd, 1983), (Palermo) Italian (Grice, 1992), Japanese (Pierrehumbert & Beckman, 1988), Serbo-Croat (Inkelas & Zec, 1988) and Swedish (Bruce, 1977) — although the pitch accents and prosodic structures identified differ somewhat for each language. With this extensive multi-lingual support, it seems reasonable to adopt a tone sequence type of description of intonation in preference to pitch movement configurations in considering language teaching. It appears that the intonation of many languages can be more readily analysed in terms of high and low

tones than in terms of pitch movements. Thus it can be assumed that the intonation of English spoken by a non-native speaker (coloured by mother-tongue interference) can, perhaps, be more robustly described by tone sequences and that the problems associated with descriptions based on pitch movement configurations can be avoided.

Such a description is particularly suited to the phonetically motivated basis for the work presented in this thesis, since in the configuration theory, the decision to recognise a feature of a pitch contour as a phonologically significant event is based on the observation that its use has a semantically contrastive function. Tone sequence theory, however, tends towards a phonetic specification of intonation; leaving the semantic function to be described by a higher level of analysis. The prospect that the language dependent linguistic function of intonation can be isolated from the intonational description is a possible contributing factor as to why the intonation of so many languages has been described in terms of high and low tone sequences. In describing intonation to a foreign learner in terms of semantically contrastive categories, a language dependent approach must be followed. However, in order to analyse the student's non-native intonation (which, in the worst case, could be the intonation of the student's native language) a language independent approach must be adopted. In taking a language independent approach, phonetic differences between the intonation of the student's speech and the intonation of a native's speech (for the same utterance) may be detectable which, however, are *not* semantically distinct.

2.2.4 Accented and stressed syllables: Terminology

At this point it is important to clarify the concepts of an *accented* syllable and of a *stressed* syllable. Section 2.1.1 referred to the phenomenon of sentential stress as the relative prominence of syllables in connected speech, and this thesis has, until now, referred to syllables exhibiting this phenomenon as being *prominent*. Prominent syllables are subdivided into accented syllables and stressed (non-accented) syllables.

The view taken by many researchers (after Bolinger, 1958) has been that pitch movement is the fundamental element of syllable prominence. That is to say, pitch movements are taken as the essence of the phenomenon of stress. Beckman (1986) has revived a traditional distinction between, what she calls, "stress accent" and "non-stress accent."

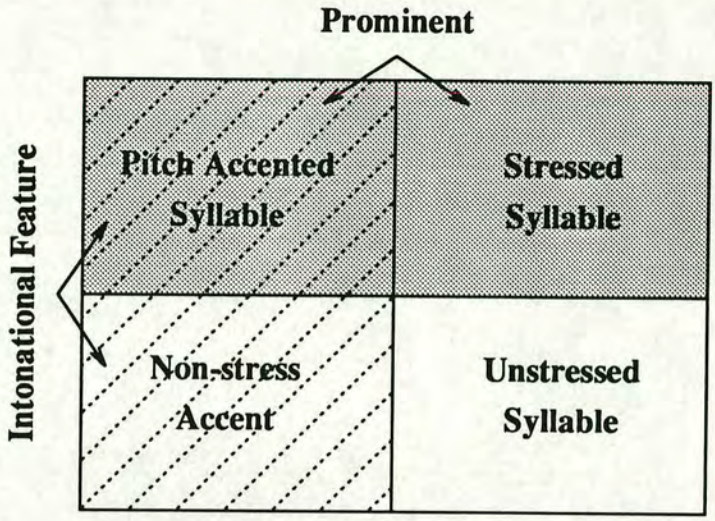


Figure 2.5: Prominent syllables and intonation features

In Beckman's view (which is adopted here) prominent syllables are initially cued by duration, intensity, and vowel quality (and/or some other type of property related to spectral quality). Pitch accents, on the other hand, are taken as intonational features. However, as described in Section 2.2.3, pitch accents are associated with prominent syllables and therefore serve as a cue to syllable prominence. Prominent syllables which are not associated with a pitch accent are called *stressed* syllables, while those which are associated with a pitch accent are called *accented* syllables. Non-stress accent refers to those tones which are intonational features not related to prominent syllables (such as those associated with prosodic structure above the level of the prosodic word — **H–**, **L–**, **H%** and **L%**). All other syllables are simply referred to as being *unstressed*. This terminology is illustrated in Figure 2.5.

Increases in the average energy of a syllable and/or increases in the duration of an entire syllable, and increases in the peak of a syllable's fundamental frequency can be used to manifest several degrees of *emphasis* (Godfrey & Brodsky, 1986). This should not be confused with the degrees of sentential stress (see Section 2.2.1) although they are closely related. The degrees of emphasis referred to by Godfrey & Brodsky are a continuum of phonetic realisations of emphatic stress in pitch accented syllables. Degrees of emphasis

are not associated with (unaccented) stressed syllables or unstressed syllables. It is assumed in this thesis that such variations in pitch range are primarily paralinguistic and beyond the scope of phonological analysis.

2.3 Discussion

It is evident from everyday experience that an utterance is perceived as being composed of a series of (lexical) words and that there is a linguistic significance to the order in which the words are concatenated. When studying a foreign language, a student does not just learn to interpret single isolated lexical items. A student must also learn the relations that can exist between words, by studying the grammar of the language. However, a student is usually not taught a formalised grammar of a language[6]. Instead, a student can learn aspects of grammar implicitly through exposure to carefully selected, illustrative examples and by attempting to mimic grammatical structures presented to the student.

It is also evident from the aforementioned studies on intonational phonology, that an utterance is made up of prosodic words (as well as lexical words) that have a linguistic significance. In teaching the prosody of a language, it should be possible to apply the same paradigm as used in teaching grammatical structures. Thus, the aim of teaching prosody is *not* to explicitly teach a formalised phonology of prosody. Furthermore, it is inappropriate to assume that a foreign language student is not naïve of linguistic theory or that the student should be made aware of abstract phonological theories which form the basis of a course syllabus. This does not, however, mean that prosodic theories can be discarded. If the same teaching paradigm is to be used in teaching prosodic structures as for teaching grammatical structures, then a knowledge of prosody must be used (say, by subscribing to a prosodic theory) to design courseware by which a student can learn stress patterns and intonation.

The theories of intonational phonology remain controversial and are in a field were there is much more research to be undertaken. The theories aim to describe a minimal set of pitch variations which are semantically contrastive. It is not clear what phonologically distinct categories a pitch contour can exhibit, but there is consensus that some

[6]There is an enormous area of research often referred to as "Natural Language Processing" or "Computational Linguistics" which is concerned with trying to formalise the grammar of languages.

patterns exist in the melodies of pitch. The prosodic structure of an utterance will convey some meaning. However, the meaning conveyed is dependent on the context of the utterance. Therefore, in teaching prosody as part of a foreign language syllabus, it is not possible to teach a student to use a single intonation contour to convey one specific meaning. Intonation must be taught in the context of well-structured dialogues. The courseware must be designed to provide such dialogues whilst subscribing to some underlying prosodic theory.

The prosodic structure of a student's speech must be compared with that of a native speaker for assessment purposes, at each stage of the dialogue. Ideally, the prosodic structure of a student's speech and a native's speech should be represented within a phonological model. Thus, the comparison would allow a student's pitch to differ from that of a native speaker only at times where it does not carry any semantically contrastive function. A corrective diagnosis could then be offered to a student in order to aid the student's pronunciation.

However, phonological representations of prosodic structure are language-specific. The prosodic structure of a student's speech can be highly influenced by mother-tongue interference and by prosodic stereotypes of the target language. This renders a language-specific phonological representation of prosodic structure inappropriate. In order to determine if a student can speak a foreign language with a prosody which is unambiguous to a native listener, it is proposed in this thesis that it is not necessary to determine if a phonological representation of the prosodic aspects of a student's speech is identical to that of a native speaker. A comparison of the prosodic aspects of speech at a phonetic level should suffice. It is therefore assumed that a higher level of analysis can be used to determine which of the detectable phonetic differences do have and do not have a semantically contrastive function in a specific language.

The research presented in this thesis aims to automatically generate a phonetic representation of the prosodic aspects of speech from an acoustic waveform. The resultant phonetic representation allows the comparison of the stress and intonation of a student's speech with that of a native speaker.

Chapter 3

Acoustic-Prosodic Analysis: An overview of related work

The automatic analysis of speech in computer aided pronunciation teaching uses a digitally sampled acoustic waveform as the only input parameter. An acoustic waveform can be easily captured as an electronic signal with a microphone without being cumbersome or obtrusive to a speaker. Other characteristic aspects of speech, such as glottal activity, nasality and air flow, cannot be measured without using cumbersome (sparsely available) transducers such as a laryngograph which ties around the neck, a nasograph which pinches below the bridge of the nose, or an obtrusive face mask.

The acoustic speech signal conveys linguistic and paralinguistic information. The acoustic waveform is perceived as a sequence of phonemic segments which constitute the words of an utterance. It also exhibits suprasegmental phenomena which are related (but not exclusively) to the syntactic and prosodic structures of an utterance. It is emphasised, therefore, that the acoustic waveform is an extremely complex signal which conveys many aspects of speech other than prosody alone. In order to perform automatic prosodic analysis of speech, it is necessary to identify acoustic parameters which are correlated with prosody. Research conducted to identify acoustic correlates of stress is reviewed in Section 3.1.

The fundamental frequency of an acoustic speech signal is correlated with both the perceived prominence of syllables and the characteristic melody of an utterance. Its extraction from the acoustic waveform is therefore a prerequisite for prosodic analysis. The functionalities of a selection of fundamental frequency determination algorithms are reviewed in detail in Section 3.2.

Prosodic aspects of speech are described in a syllabic domain. Thus, the automatic identification of syllables in connected speech forms a part of prosodic analysis. A number of algorithms to partition speech into syllabic units are reviewed in Section 3.3.

It is asserted that the acoustic correlates of stress are influenced by factors other than stress and interact in the manifestation of stress. The fundamental frequency of speech is therefore not a direct representation of intonation and acoustic correlates of stress do not co-exist in a simple relationship to represent stress. The composite structure of fundamental frequency is discussed in Section 3.4. Algorithms proposed to automatically transcribe the prosodic structure of an utterance from acoustic parameters are reviewed in Section 3.5.

3.1 Acoustic parameters for prosodic analysis

Extensive research has been conducted to identify the acoustic correlates of lexical stress in isolated words and of sentential stress in connected speech.

Fry (1955) examines acoustic features correlated with a shift in lexical stress from one syllable to another which is commonly associated with a change in function from noun to verb in disyllabic words in English such as *digest* and *permit*. Fry's results show that the location of the lexical stress is correlated with the ratio of the duration of the vowel in the syllables and the ratio of the peak of energy in the syllables (which occurred in the vocalic part of the syllables). He also claims that the duration feature is a more effective cue to stress location than the energy feature. In a later experiment, Fry (1958) investigates the correlation of fundamental frequency with stress shift in the same type of words. He finds that the syllable with the greatest fundamental frequency is more likely to be perceived as prominent although the extent of the jump in fundamental frequency ($F\emptyset$) between the two syllables is relatively unimportant. Moreover, Fry suggests that the occurrence of an $F\emptyset$ glide through a syllable is likely to make it perceptually more salient than the other syllable (in the disyllabic words he examines) and that such glides may "outweigh" the duration cue.

Fry also draws attention to a further correlate of stress, vowel quality.

> "[I]n English ...[t]he substitution of the neutral vowel /ə/ for some other vowel, the reduction of a diphthong to a pure vowel, or the centralisation of

a vowel are all powerful cues in the judgement of stress."

(Fry, 1958, pp.128)

The effect of vowel quality on the perception of lexical stress is also investigated by Fry (1965). The experiment he conducted examines the relation between stress location and changes in the *F1/F2* formant space of monophthongs using synthetic stimuli. Although Fry finds that stress judgement is dependent on vowel formant structure, he suggests that the perception of stress is affected more by duration and energy cues than by vowel quality.

The four acoustic features associated with the perception of stress (identified by Fry as duration, energy, fundamental frequency and vowel quality) have been the subject of investigation for many researches (Lieberman, 1960; Morton & Jassem, 1965; Westin *et al.*, 1966; Nakatani & Aston, 1978; Aull & Zue, 1985).

Lieberman (1960) proposes an algorithm to identify the location of the lexical stress in minimal noun/verb pairs of disyllabic words such as '*compound* and *com'pound*, using duration, energy and fundamental frequency (but not vowel quality). Lieberman suggests that the duration of the entire syllable (cf. the duration of the vowel used by Fry) acting as a unidimensional cue for the location of the most prominent syllable is a less effective cue than the peak syllable amplitude. This is the converse to the claim of Fry (1955). Lieberman also suggests that the peak syllable fundamental frequency is a more effective cue to stress location than the peak syllable amplitude. Lieberman recognises the complex interaction that exists between these acoustic features. In some cases, the prominent syllable has a higher peak *FØ*, but the unstressed syllable has a higher peak amplitude; and in other cases, the prominent syllable has a higher peak amplitude, but the unstressed syllable has a higher peak *FØ*. This is analogous to Fry's observation that *FØ* glide through a syllable may "outweigh" the duration cue. In other words, *FØ* may also "outweigh" the energy cue, or visa versa. By combining these acoustic features, Lieberman is able to correctly identify the location of the lexical stress in 99.2% of the minimal noun/verb pairs he studies.

Adams & Munro (1978) examine the acoustic correlates of sentential stress (in connected speech). Their investigation considers fundamental frequency during the course of syllables, the absolute duration of entire syllables and the amplitude envelope in sylla-

bles. Adams & Munro do not consider the effects of vowel quality mentioned by Fry. The results of Adams & Munro suggest that the degree of *FØ* change is a more significant correlate of sentential stress than absolute *FØ* height.

> "[T]he initial level and the end level were higher in type I [*rising FØ*] and lower in type II [*falling FØ*] for stressed than for unstressed syllables. Furthermore, the amount of rise in type I and fall in type II was found to be greater in stressed than in unstressed syllables. ... In type V [*level FØ*], the fundamental frequency level was greater in stressed than in unstressed syllables, but *not significantly so*." [my italics]
>
> (Adams & Munro, 1978, pp.140)

The results of Adams & Munro also indicate that the amount of fall from a steady peak in the amplitude envelope is a more significant correlate of sentential stress than the amplitude at the peak of the envelope, and that sentential stress is also signalled by syllable duration.

There is noticeably little reference to vowel quality in these studies on the perception of stress. Most of the work relating stress with vowel quality is found in studies on the perceptual classification of vowels. Such studies take a different view of the relation between vowel quality and stress to the view taken in this thesis, in that they are primarily concerned with understanding how stress, and other factors, affect vowel quality; whereas here an emphasis is placed on understanding how vowel quality, and other factors, affect the perception of stress.

Tiffany (1959) observes a tendency for the vowel diagram (a conventional plot of the first formant *F1* against the second formant *F2*) to grow smaller from vowels enunciated in isolation to prominent vowels to non-prominent vowels. The prominent and non-prominent vowels which Tiffany investigates are in a [h]-vowel-[d] phonetic context embedded in carrier phrases with and without emphatic stress placed on the word containing the relevant vowel. Tiffany suggests that vowels move towards a neutral, or at least a central, point on the vowel diagram as they lose energy. Prominent vowels may not merely be longer in duration, higher in fundamental frequency, and greater in energy than unstressed vowels; they may also be different in terms of vowel resonance patterns as well.

Shearme & Holmes (1962) observe that *F1/F2* measurements for vowels in connected

speech and in varying phonetic contexts are displaced from corresponding measurements for vowels in isolated monosyllables towards a central vowel position. They attribute this affect to variability in both stress *and* phonetic context.

It is argued by Lindblom (1963) that as rate of articulation increases, physical limitations impose less time for the articulators to complete their movements within a consonant-vowel-consonant syllable, thus the vowel becomes shorter and hence the vowel becomes more susceptible to contextual assimilation. The acoustic consequence of this is that the placement of the vowel in the *F1/F2* formant space moves away from a target position, typically towards a more central vowel position. Lindblom argues that vowel reduction is primarily determined by a duration feature and that it is immaterial to discuss whether the duration of a vowel is produced chiefly by the degree of stress or the rate of articulation. If Lindblom's explanation of vowel reduction in terms of a shortening of duration and the physical limitations of the articulators, rather than variations in the degree of stress and the rate of articulation *per se*, is correct, then it is hypothesised that duration acts as a stronger cue to stress than vowel quality and that the correlation between duration and vowel quality is positive in their association with stress. In other words, no more information about the degree of stress of a syllable will be obtained from a measure of vowel quality than can already be obtained from a duration feature.

Moreover, if Lindblom's explanation is correct then it is inferred that variations in the degree of stress and the rate of articulation affect the vowel quality in identical ways. Verbrugge & Shankweiler (1977) report that measurements of vowel formant frequencies reveal relatively little difference between fast and slow speech, but reveal large vowel formant shifts in unstressed syllables relative to stressed syllables. Tuller *et al.* (1982) examine whether stress and speech rate variations involve different transformations of physiological signals. The hypothesis that all changes in vowel duration are the product of the same production rule, as argued by Lindblom, is not supported by their results. Variations in stress and speech rate apparently produced vowel duration changes through different effects on muscle (genioglossus and orbicularis oris) activity. Evidence against Lindblom's hypothesis is also presented by van Summers (1987) who argues that duration lengthening due to final consonant voicing in consonant-vowel-consonant syllables has greater effects on articulatory movements and formant structure in the final portions of

vowels than on the initial portions, while duration lengthening due to increased syllable prominence has a more global influence throughout vowel production.

Vowel reduction is associated with the degree of stress and the rate of articulation (Lindblom, 1963; Tuller *et al.*, 1982; van Summers, 1987). Therefore, if vowel quality is to function as a cue to stress location, it is first necessary to understand how vowel quality is affected by factors other than stress. If the reduction in the quality of a vowel can be attributed to variations in the rate of articulation, for example, rather than variations in the degree of stress then there is little possibility of being able to use vowel quality as a cue to stress.

The consensus drawn from these studies is that duration, fundamental frequency, energy and the quality of the syllable nucleus correlate with lexical and sentential stress, although the degree to which each of these parameters correlates with the stress varies considerably from study to study. One of the major problems in the automatic analysis of stress is that there are complex interactions between each of these acoustic parameters in their association with stress, and variations in each of these parameters can be associated with factors other than stress.

For example, the duration of the vocalic portion of a word is influenced by the inherent or intrinsic duration of the intended vowel, phonetic context (Lehiste, 1970), rate of articulation, the sentential stress pattern, the position of the word within the sentence, and other possible factors such as speaker style. Although duration may be ambiguous as an unidimensional cue to stress, syllable final consonant voicing, or any of the other factors affecting duration, it may provide useful information for, say, stress when used in combination with other acoustic cues. If the presence of stress increases vowel duration along with having some specific influence on formant trajectories which final consonant voicing does not, then the stress related vowel lengthening could be disambiguated from voicing related lengthening.

Lehiste & Peterson (1959) investigate intrinsic vowel amplitude and suggest that energy cues to the type of stress investigated by Fry can be more readily identified by first compensating for intrinsic vowel amplitudes. This philosophy can be extended to include any acoustic cue to stress. If acoustic features are normalised for variations which are due to factors other than stress, then the normalised acoustic features can be

combined to form a description of the stress pattern of an utterance.

In summary, the phenomenon of stress is realised in the speech waveform by variations in duration, energy, fundamental frequency and vowel quality. The process of locating prominent syllables is complicated both by the interaction of these features in the manifestation of stress and by the fact that all four features are also influenced by factors other than stress. Moreover, there are a number of specific problems in measuring these acoustic features in order to automatically locate prominent syllables. It is not clear how the acoustic features are determined, over what phonetic domains the acoustic features are related with respect to prosody, or how they are best normalised for non-prosodic aspects of speech. A principal aim of this thesis is to address these problems.

Methods to automatically determine the fundamental frequency of a speech waveform are addressed in Section 3.2 and in Chapter 7. The domain of phonetic units whose duration, energy and vowel quality are to be determined as optimal correlates of stress and normalisation techniques applied to them are investigated in Chapters 4, 5 & 6.

3.2 Automatic extraction of fundamental frequency

The *fundamental frequency* (*FØ*) of speech is defined as the rate of glottal pulses generated by the vibration of the vocal folds during the voicing of segments. The *pitch* of speech is the perceptual correlate of *FØ*. The psychoacoustic scales of pitch are linear only at relatively low frequencies. However, it is assumed that there is a linear correlation between pitch and *FØ* at the low ranges of frequency that are relevant to the voicing of male and female speech (approximately 50–250Hz and 120–400Hz respectively).

It can be concluded from the previous Sections that the fundamental frequency of speech plays an important role in the prosodic features of stress and intonation. Its extraction from the acoustic speech waveform is therefore necessary as an initial process for the analysis of these suprasegmental phenomena. However, determining *FØ* is not a simple task, and many approaches (referred to here as Fundamental frequency Determination Algorithms (FDAs)[1]) have been reported (Hess, 1983). The diversity

[1]Methods of extracting fundamental frequency are commonly referred to in the literature as Pitch

and complexity of methods used in an attempt to determine *FØ* stem from the non-stationarity of speech, characterised by non-uniform intensity, by small variations in fundamental frequency between successive periods (such as those particularly noticeable in creaky voice (Laver, 1980)) and by a constantly changing spectrum which is dependent upon articulation.

A selection of FDAs are described below. The choice of FDAs reviewed here is influenced by the availability of existing implementations, by the desire to examine methods of fundamental frequency extraction which use radically different techniques, and by the ease of implementation from the original descriptions of the algorithms. The algorithms investigated are:

- Cepstrum-based *FØ* determinator (CFD) (Noll, 1967).
- Harmonic product spectrum (HPS) (Schroeder, 1968; Noll, 1970).
- Feature-based *FØ* tracker (FBFT) (Phillips, 1985).
- Parallel processing method (PP) (Gold & Rabiner, 1969).
- Integrated *FØ* tracking algorithm (IFTA) (Secrest & Doddington, 1983).
- Super resolution *FØ* determinator (SRFD) (Medan *et al.*, 1991).

The algorithms CFD and HPS make use of frequency domain representations of the speech signal. FBFT and PP produce fundamental frequency estimates by analysing the waveform in the time domain. IFTA and SRFD uses a waveform similarity metric based on a normalised crosscorrelation coefficient. This selection of FDAs is considered to represent a cross-section of the multitude of algorithms developed over recent years. A detailed performance evaluation of these FDAs is described in Section 7.2. The remainder of this Section provides a summary of the operation of each of these algorithms in order to illustrate their diversity and complexity.

Determination Algorithms. However, these algorithms do not determine the linguistic phenomenon of pitch, which is the perceptual correlate of *FØ*.

3.2.1 Cepstrum-based *FØ* determinator (CFD)

Noll (1967) proposes a method to determine the fundamental frequency of a speech waveform based on the use of *cepstra*. A cepstrum is defined as the Fourier transform (FFT) of the logarithm power spectrum of a signal. The method assumes that a speech signal can be represented by the commonly understood source-filter model (Fallside & Woods, 1985) in which the speech signal $f(t)$ is equal to the convolution of the vocal source signal $s(t)$ and the impulse response of the vocal tract $h(t)$.

$$f(t) = s(t) \underset{\omega}{\otimes} h(t) \tag{3.1}$$

The objective of the algorithm is to effectively deconvolve the vocal tract response from the source signal, and thus find the fundamental frequency of the speech.

Following from Equation 3.1, the source-filter model can be represented in the frequency domain (by application of the convolution theorem) as,

$$F(\omega) = S(\omega).H(\omega) \tag{3.2}$$

In order to identify or separate the effects of the vocal tract and the vocal source signal, a Fourier transform is performed on the logarithm of the power spectrum, $|F(\omega)|^2$,

$$\begin{aligned} \mathcal{FFT}\left\{\log|F(\omega)|^2\right\} &= \mathcal{FFT}\left\{\log(|S(\omega)|^2.|H(\omega)|^2)\right\} \\ &= \mathcal{FFT}\left\{\log|S(\omega)|^2 + \log|H(\omega)|^2\right\} \\ &= \mathcal{FFT}\left\{\log|S(\omega)|^2\right\} + \mathcal{FFT}\left\{\log|H(\omega)|^2\right\} \end{aligned} \tag{3.3}$$

Thus, the vocal source and vocal tract effects are now added rather than convolved. The effect of the vocal tract appears in the cepstrum as numerous, closely packed peaks at the lower quefrency[2] end, and the location of a peak occurring at the higher end of the cepstrum marks the fundamental period of the vocal source.

An illustration of the practical application of this method is shown in Figure 3.1. Cepstral analysis is performed on successive frames of data. The duration of the analysis

[2]The quefrency refers to the frequency of the ripples in a spectrum. Since the ripples are in the frequency domain, quefrency is in the time domain.

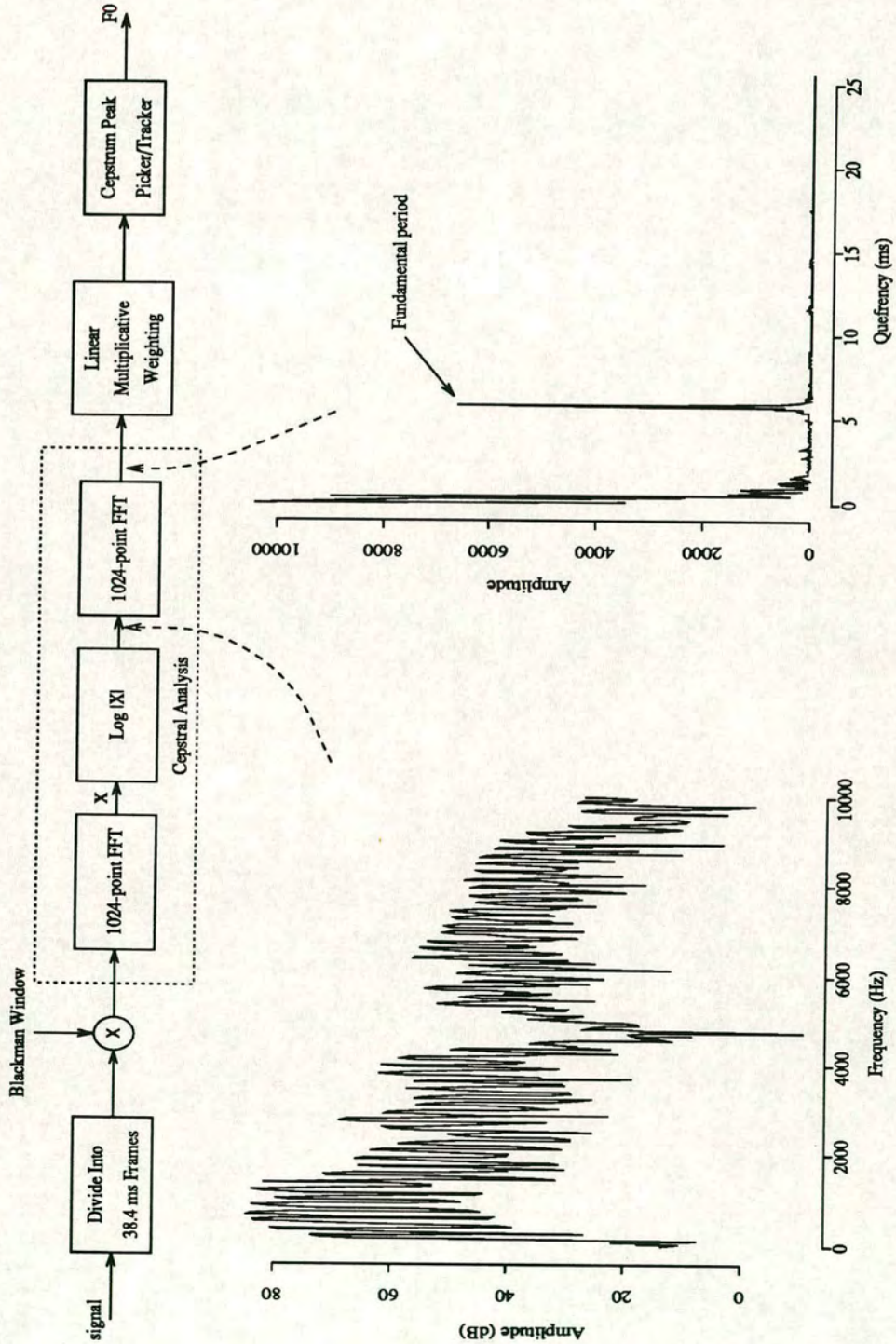


Figure 3.1: Cepstrum-based *FØ* determinator (CFD)

frame must be chosen with care. If the duration is too long, then the non-stationarity requirement of the analysis frame for the Fourier transform will be violated, particularly during segmental transitions, and spurious artefacts will be produced in the spectra. However, if the duration is too short, then the analysis will be inaccurate. In practice, a duration of 38.4ms is used as a reasonable compromise, as it corresponds to $\frac{3}{4}$ of the data in a 1024-point Fourier transform for data sampled at 20kHz, and corresponds to approximately two fundamental periods at the lower levels of *FØ* for male speech.

Cepstral analyses of sections of a speech waveform are overlapped by spacing the beginnings of successive analysis frames at intervals of 6.4ms ($\frac{1}{8}$ of the FFT data). An estimate of the fundamental frequency at the time which corresponds to the mid-point of the analysis frame is therefore made at regular intervals. The *FØ* estimates are plotted against these times to form an *FØ contour*.

Each analysis frame is applied through a Blackman window (Harris, 1978) and a 1024-point Fourier transform is performed on the data with padded zeros. The use of a 1024-point FFT results in a cepstrum of 512 coefficients (0–25.6ms) with a quefrency resolution of $\frac{1}{20}$ms (for a speech waveform sampled at a frequency of 20kHz). The quantisation error for each estimation of the fundamental period is therefore $\frac{1}{20}$ms. After cepstral analysis of the data, a linear multiplicative weighting is applied over a speaker-dependent quefrency range. For example, a quefrency range of 4.0ms (250Hz) to 20.0ms (50Hz) is selected for the analysis of male speech, and a range of 2.5ms (400Hz) to 8.3ms (120Hz) is selected for female speech. Noll (1967) proposes that the amplitude of the cepstral coefficients is weighted by 1.0 at the lower quefrency limit and on a linear scale up to a weight of 5.0 at the upper quefrency limit. The peak value in the resultant weighted cepstrum is located over the selected quefrency range. If the amplitude of the peak exceeds a speaker-dependent *a priori* threshold, then the peak location corresponds to the fundamental period for that frame; otherwise the frame is assumed to represent unvoiced speech.

3.2.2 Harmonic product spectrum (HPS)

The fundamental frequency of a periodic signal can be determined by measuring the frequencies of its higher harmonic components and computing the greatest common divisor

of these harmonic frequencies (Schroeder, 1968). The greatest common divisor can be determined by making an entry to a frequency histogram for each harmonic frequency and at integer divisions of the harmonic frequency. The frequency at the peak of the histogram represents the greatest common divisor of the harmonic frequencies, and hence the fundamental frequency. However, the task of determining the harmonic frequencies is not simple. The problem of determining these frequencies can be avoided by noting that the frequency components of a signal represented in the frequency domain have higher amplitudes at points near to harmonic frequencies. Each entry to the histogram can thus be weighted by a monotonicly increasing function of the component amplitude or of the logarithm of the component amplitude.

The block diagram shown in Figure 3.2 illustrates how the HPS based *FØ* determinator is arranged in practice. The speech signal is divided into 38.4ms frames with an analysis frame shift of 6.4ms, as in the cepstral based approach (CFD). A 4-term Blackman-Harris window (Harris, 1978) is applied to each frame in order to filter out discontinuities (and hence high frequency artefacts) at the limits of the analysis frame. The harmonic histogram is formed from a short term log power spectrum $20\log_{10}|F(nf)|$ and represented as the 'harmonic product spectrum' (Noll, 1970),

$$\log_{10} P(f) = \sum_{n=1}^{N} 20.\log_{10}|F(nf)| \tag{3.4}$$

$$P(f) = \prod_{n=1}^{N} |F(nf)| \quad (+10^{20}, \text{which is ignored}) \tag{3.5}$$

The low-quefrency structure of the log power spectrum of speech only adds confusion when compressed to form the harmonic product spectrum, so such terms are removed by smoothing the spectrum with a window function $W(f)$ (for example, a 4-term Blackman-Harris window) and then subtracting the smoothed spectrum from the original spectrum prior to calculating the harmonic product spectrum. The low-quefrency lifted log spectrum is expressed as,

$$L(f) = 20\log_{10}|F(f)| - W(f).20\log_{10}|F(f)| \tag{3.6}$$

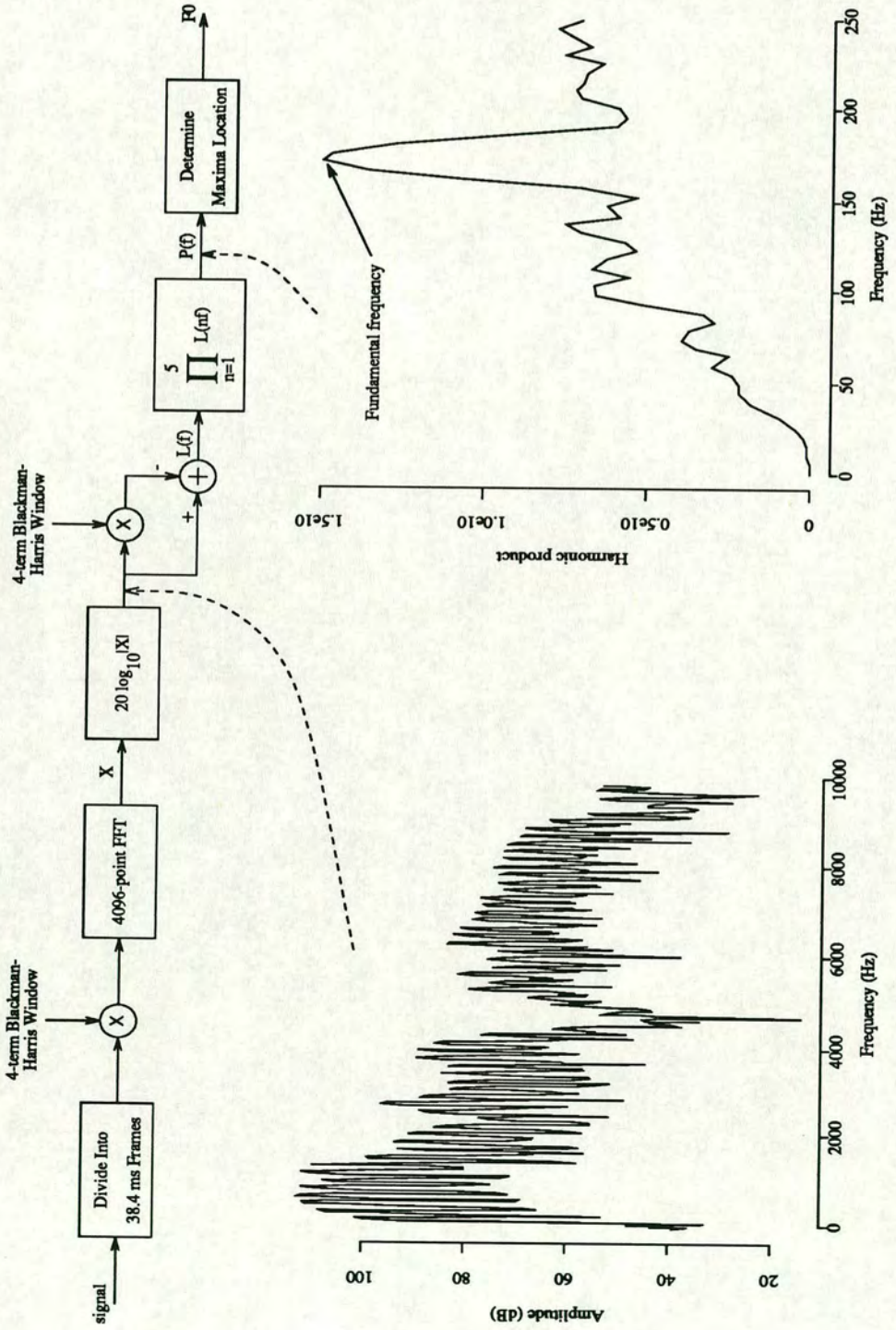


Figure 3.2: Harmonic product spectrum (HPS)

and the harmonic product spectrum, as,

$$P(f) = \prod_{n=1}^{N} L(nf) \tag{3.7}$$

where N is a compression factor (set to 5 in a practical application). If the compression factor is set too low, the estimates of *FØ* will be unreliable, and if the compression factor is set too high, the process of determining *FØ* by this method will be time consuming and computationally expensive.

The fundamental frequency is given by the position of the maxima of $P(f)$, which is searched for in the range of 50Hz to 250Hz for male speech, and between 120Hz and 400Hz for female speech (for compatibility with the other FDAs reviewed in this Section). If this maxima has an amplitude that is greater than some speaker-dependent *a priori* threshold, then the section of speech being analysed is assumed to be voiced. For a speech waveform sampled at 20kHz and with the use of a 4096-point FFT, the fundamental frequency can be estimated by this method with a quantisation error of 4.88Hz (ie. the sampling frequency divided by the number of points in the FFT).

3.2.3 Feature-based *FØ* tracker (FBFT)

An algorithm is proposed by Phillips (1985) which aims to locate glottal pulses in the time domain by using a feature-based statistical approach. The algorithm uses perceptually motivated features that are designed to capture the same information that a person would use to indicate the presence of a glottal pulse from a waveform display.

As is the case with many FDAs which operate on a time domain representation of speech (such as those reviewed in this and the following Sections), the sampled waveform is initially low-pass filtered to simplify its temporal structure. The filtering process reduces the effects of higher formants (*F2*, *F3*, and *F4*) for vowels, and removes the high frequency components of voiced fricatives, thus aiding the classification of voiced speech by the algorithms. In the studies reported here, low-pass filtered speech is produced by a finite impulse response (FIR) filter with a -3dB cut-off at 600Hz and rejection greater than -85dB above 700Hz, though these filter characteristics are not crucial to the performance of the FDAs.

Peaks are located in the speech waveform under the criteria that they correspond to local maxima with signal amplitudes greater than 120 ADC-units[3] (approximately 50dB signal-to-noise level) and that the signal amplitude drops by at least 20% after the local maxima. Peaks satisfying these criteria are represented by the set S_1 and are indicated on the speech waveform in Figure 3.3 by ▾and ▼.

A member of S_1 is accepted into a subset, S_2, if no other peak occurs within the 3ms region before it, or if its amplitude is greater than 75% of the amplitude of the largest peak within the 3ms region before it. The members of the subset S_2 are indicated in Figure 3.3 by ▼.

Perceptually motivated features are used to identify the first peak of each fundamental period in the speech waveform during voiced portions of speech. The peaks identified in S_2 are considered as candidates for the first peak of a fundamental period. The following nineteen measurements $\{M_i \mid i \in 1, \ldots, 19\}$ are made for pairs of candidates, $P_1, P_2 \in S_2$ (illustrated in Figure 3.3); P_i at time t_i with amplitude v_i:

- $M_1 = v_1$; the waveform amplitude at peak P_1.
- $M_2 = (v_2 - v_1)/v_1$; the relative change in amplitude between the candidate peaks.
- $M_3 = t_2 - t_1$; the duration between the candidate peaks.
- $M_4 = w_1$; the width at the zero crossing point of the peak P_1.
- Locate the largest peak $P_0 \in S_1$ within the region of duration $1.5(t_2 - t_1)$ before P_1. $M_5 = v_0/\min(v_1, v_2)$; the amplitude of peak P_0 relative to the minimum of the amplitudes of the candidate peaks.
- Locate the largest peak $P_3 \in S_1$ within the region of duration $1.5(t_2 - t_1)$ after P_2. $M_6 = v_3/\min(v_1, v_2)$.
- $M_7 = (v_3 - v_2)/v_2$.
- $M_8 = (t_3 - t_2)/(t_2 - t_1)$.
- Locate the largest peak $P_a \in S_1$ within the 3ms region before P_1. $M_9 = (v_1 - v_a)/v_1$.

[3]Speech signals are quantised by a 16-bit analogue-to-digital converter (ADC).

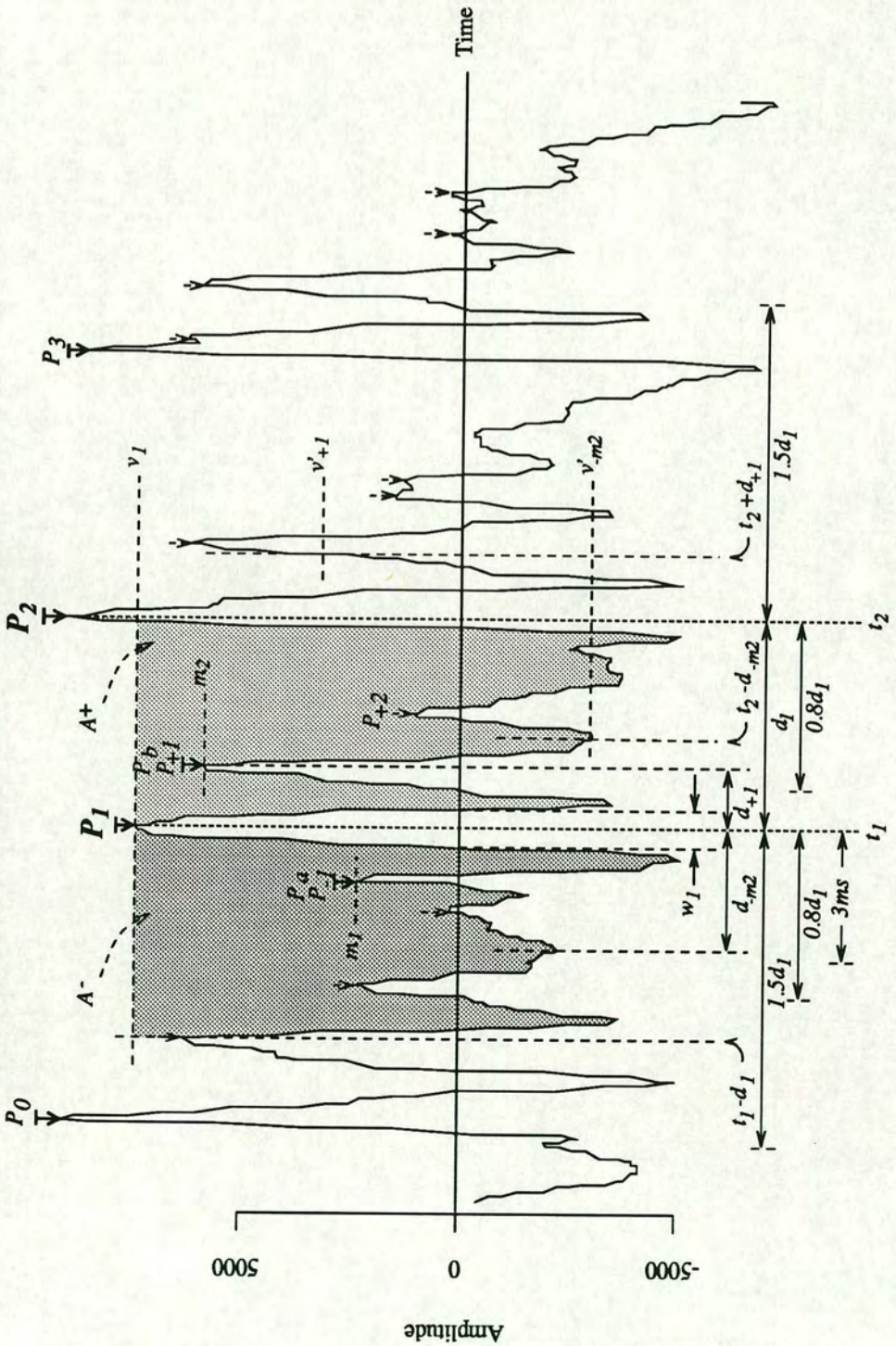


Figure 3.3: Feature-based *FØ* tracker (FBFT)

- Locate the largest peak $P_b \in S_1$ that lies between the two candidate peaks. $M_{10} = v_b / \min(v_1, v_2)$.
- Locate the first peak $P_{+1} \in S_1$ after P_1 and determine the waveform amplitude v'_{+1} at the same time after P_2. $M_{11} = (v'_{+1} - v_{+1})/v_1$.
- Locate the second peak $P_{+2} \in S_1$ after P_1 and determine the waveform amplitude v'_{+2} at the same time after P_2. $M_{12} = (v'_{+2} - v_{+2})/v_1$.
- Locate the first peak $P_{-1} \in S_1$ before P_1 and determine the waveform amplitude v'_{-1} at the same time before P_2. $M_{13} = (v'_{-1} - v_{-1})/v_1$.
- Locate the first local minimum after P_1 and determine the waveform amplitude v'_{+m1} at the same time after P_2. v_{+m1} is the amplitude of the first local minimum after P_1. $M_{14} = (v'_{+m1} - v_{+m1})/v_1$.
- Locate the first local minimum before P_1 and determine the waveform amplitude v'_{-m1} at the same time before P_2. v_{-m1} is the amplitude of the first local minimum before P_1. $M_{15} = (v'_{-m1} - v_{-m1})/v_1$.
- Locate the second local minimum before P_1 and determine the waveform amplitude v'_{-m2} at the same time before P_2. v_{-m2} is the amplitude of the second local minimum before P_1. $M_{16} = (v'_{-m2} - v_{-m2})/v_1$.
- Determine the maximum waveform amplitude, m_{-1} over the region t_1 to $t_1 - 0.8(t_2 - t_1)$. $M_{17} = (v_1 - m_{-1})/v_1$.
- Determine the maximum waveform amplitude, m_{-2} over the region t_2 to $t_2 - 0.8(t_2 - t_1)$. $M_{18} = (v_2 - m_{-2})/v_2$.
- The area A^+ is defined as the sum of differences in waveform amplitude and the amplitude of the peak P_1 over the region t_1 to $t_1 + (t_2 - t_1)$. The area A^- is defined as the sum of differences in waveform amplitude and the amplitude of P_1 over the region t_1 to $t_1 - (t_2 - t_1)$. $M_{19} = A^-/A^+$.

These measurements are combined in a statistical classifier that derives the location of glottal pulses in the waveform. The set of glottal pulse times generated by this method

are considered in chronological order. The duration between adjacent glottal pulses is calculated and converted to Hertz. If the value lies within a limited range, it is taken to represent the fundamental frequency at the time located between the two glottal pulses; otherwise, the duration between the marks is considered to correspond to an unvoiced region of speech. The *FØ* values are limited to lie between 50Hz and 250Hz for male speech, and between 120Hz and 400Hz for female speech.

3.2.4 Parallel processing method (PP)

The parallel processing approach to *FØ* extraction proposed by Gold & Rabiner (1969) uses multiple peak picking in the time domain. An outline of this method is shown in Figure 3.4. The speech signal is initially low-pass filtered to simplify the temporal structure of the sampled waveform, as in the feature-based algorithm (FBFT). In order to make an unbiased comparison of the algorithms being reviewed here, it is necessary to set parameters which are common across algorithms to the same values. Thus, the signal is divided into a series of 38.4ms duration analysis frames with successive frames being separated by a 6.4ms frame shift, as with the other FDAs. Each analysis frame is processed by a 'silence' (low energy) detector. If two or more samples in a frame have magnitudes that exceed 120 ADC-units (approximately 50dB signal-to-noise level) then the frame is assumed not to be silence, and the frame is processed through six simple peak/valley detectors, each examining a different aspect of the waveform. The six detectors each generate a pulse train in which the magnitude of each pulse is governed by the particular aspect of the waveform being measured. The six measurements made by the detectors for the local maximum (peak) and the local minimum (valley) in the filtered waveform are:

- M_1, magnitude of each local maximum (peak).
- M_2, magnitude of each local minimum (valley).
- M_3, absolute difference in the amplitude of each peak and the amplitude of the previous valley.
- M_4, absolute difference in the amplitude of each valley and the amplitude of the previous peak.

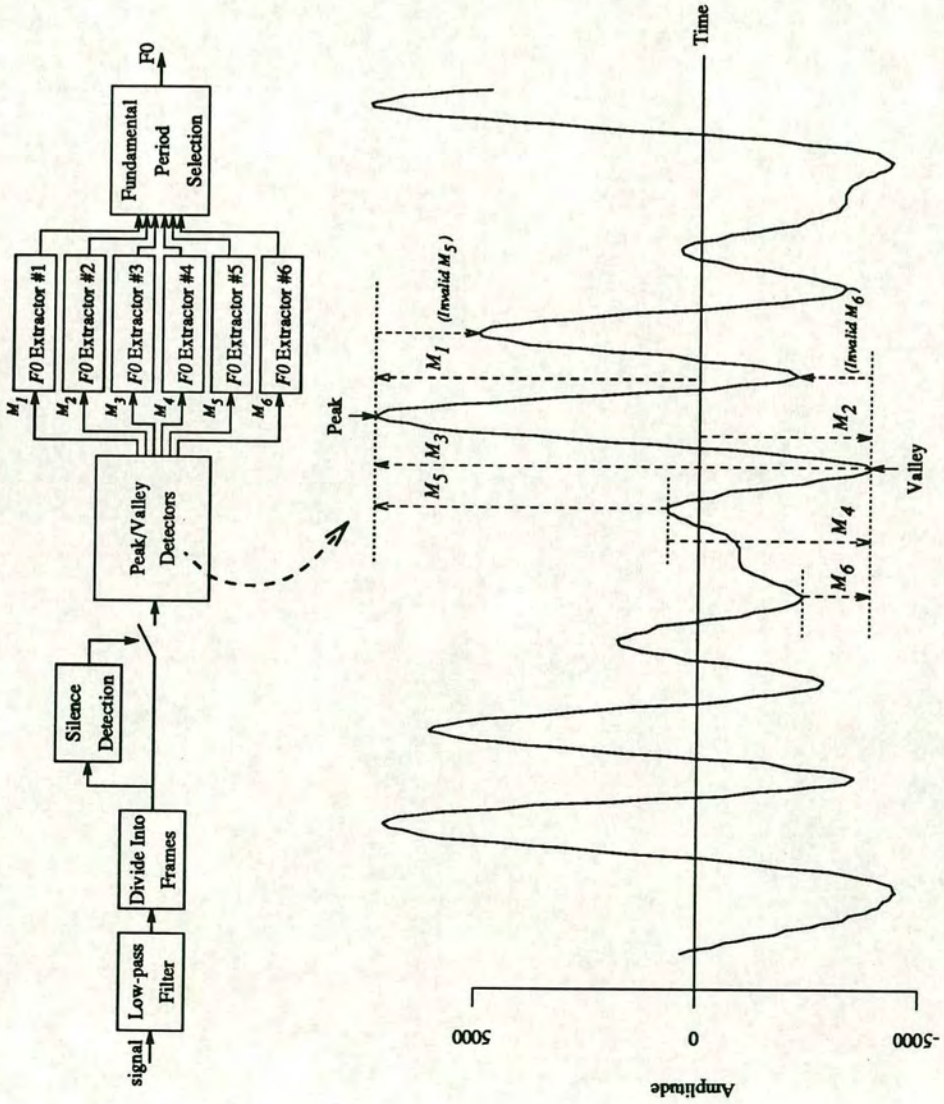


Figure 3.4: Parallel processing method (PP)

- M_5, difference in the amplitude of each peak and the amplitude of the previous peak (never permitted to be negative — ie. this measure is invalid when the peak amplitude is less than the amplitude of the previous peak).
- M_6, difference in the amplitude of each valley and the amplitude of the previous valley (never permitted to be negative – ie. this measure is invalid when the valley amplitude is greater than the amplitude of the previous valley).

The peak/valley detectors are followed by six individual *FØ* extractors operating in parallel. The *FØ* extractors work in the following way. After each detection of a pulse from the preceding peak/valley detector i, a threshold is set to the magnitude of the measurement M_i. The threshold remains constant for a blanking interval τ (during which time no subsequent pulse detection is allowed) followed by a single exponential decay (with time constant β). Whenever, the magnitude of an incoming pulse exceeds the level of the decay threshold value, it is detected and the process is repeated. The estimate of the fundamental period is the duration between each consecutive pair of detected pulses. The two parameters, τ and β, are made dependent upon a running average of the fundamental period. Finally, an overall estimate of fundamental period is computed by using a complex majority-wins type algorithm on the three most recent estimates of fundamental period from the six extractors (and estimates derived from these). A voiced/unvoiced decision is made on the basis of the degree of agreement between the six *FØ* extractors. If four or more agree, the frame being analysed is classified as voiced.

This algorithm is somewhat restricted by the fixed analysis frame length of 38.4ms in that at least two fundamental periods must reside within this duration for the *FØ* extractors to function adequately. For the frame length specified, the lowest fundamental frequency which the algorithm can be expected to extract reliably is approximately 52Hz.

3.2.5 Integrated *FØ* tracking algorithm (IFTA)

Secrest & Doddington (1983) propose that the linear prediction coding (LPC) (Markel & Gray, 1976) residual error signal mainly contains excitation information and therefore should theoretically provide the best signal from which to extract estimates of funda-

mental frequency. Inadequacies of the LPC model cause high frequency noise to be introduced into the residual signal, thus complicating the temporal structure of the signal and hindering *FØ* estimation. If the residual signal is low-pass filtered in an attempt to overcome this problem, then high frequency energy present in unvoiced regions of speech are also removed, rendering the signal suboptimal for making decisions as to whether or not the speech is voiced or unvoiced. Secrest & Doddington note that the first reflection coefficient of the LPC, K_1, is correlated with the high/low frequency energy in the signal and propose a single pole de-emphasis filter, $F(z) = 1/(1 - K_1 z^{-1})$, to low-pass filter the residual signal in voiced (low frequency energy) regions of speech and high-pass filter it in unvoiced (high frequency energy) regions.

Estimates of *FØ* are made from the de-emphasised residual error signal by quantifying the degree of similarity between two adjacent, non-overlapping sections. The signal is analysed frame-by-frame at intervals of 6.4ms (for compatibility with the other FDAs reviewed here). Each frame contains a set of samples, $s_N = \{s(i) \mid i \in 1, \ldots, N\}$, which is divided into two consecutive sections each containing a variable number of samples, n.

$$\begin{aligned} x_n &= \{x(i) = s(i) \mid i \in 1, \ldots, n\} \\ y_n &= \{y(i) = s(i+n) \mid i \in 1, \ldots, n\} \end{aligned} \tag{3.8}$$

Candidate values of the fundamental period are obtained by locating peaks in the normalised crosscorrelation coefficient given by,

$$r_{x,y}(n) = \frac{\sum_{j=1}^{n} x(j).y(j)}{\sqrt{\sum_{j=1}^{n} x(j)^2 . \sum_{j=1}^{n} y(j)^2}}, \qquad n \in N_{min}, \ldots, N_{max} \tag{3.9}$$

where N_{min} and N_{max} are the minimum and maximum expected fundamental period values (in number of samples) for a given speaker.

The amplitude of $r_{x,y}(n)$ quantifies the degree of similarity between x_n and y_n independently of variations in the signal energy. If x_n and y_n are in phase and have similar temporal structures, then $r_{x,y}(n)$ tends towards $+1$; and if x_n and y_n are out of phase or have different temporal structures, then $r_{x,y}(n)$ tends towards -1.

Secrest & Doddington propose an algorithm to determine the optimal fundamental period from the set of candidate values. The approach requires several frames of data to be analysed before a decision is made upon the optimal period and voicing classification of a frame. Every fundamental period candidate of a frame is compared with each candidate obtained for the previous frame. A penalty score is made for each comparison. The score penalises large deviations in fundamental period from one frame to the next, low values of $r_{x,y}(n)$ during voiced regions of speech and high values of $r_{x,y}(n)$ during unvoiced regions of speech, and uses a measure of the difference between the spectra of the two consecutive frames to penalise changes in the voicing classification when the spectrum is relatively unaltered from one frame to the next. For each candidate in a frame, there is a candidate in the previous frame for which this penalty score is minimal. A trajectory is formed by back-tracking from each candidate in a frame to the optimal candidate of the previous frame over a series of frames. The penalty score is cumulated at each step in the trajectory. The candidate selected as the optimal fundamental period for a frame is the one associated with the trajectory which has the minimum cumulative penalty score.

3.2.6 Super resolution *FØ* determinator (SRFD)

Medan *et al.* (1991) propose an algorithm which also uses a normalised crosscorrelation coefficient to quantify the degree of similarity between two adjacent, non-overlapping sections, as in the integrated *FØ* tracking algorithm (IFTA). The speech signal is initially low-pass filtered to simplify the temporal structure of the sampled waveform, as in the feature-based algorithm (FBFT) and the parallel processing method (PP).

Each frame of speech data is initially processed by a 'silence' (low energy) detector. The silence detector used in this algorithm differs somewhat from the silence detector used in the parallel processing method, which is not robust to d.c. offset. The minimum and maximum values of the sample sets $x_{N_{min}}$ and $y_{N_{min}}$ (see Equation 3.8) are determined. If the sum of their absolute values is less than some preset threshold (say, 240 ADC-units to be consistent with the 50dB signal-to-noise level used in PP) for either set of samples, then the current frame under analysis is classified as silence and no further analysis is done on the frame.

If the frame of data is not classified as silence, then candidate values for the fundamental period (in number of samples) are sought from values of n within the range N_{min} to N_{max} by using the normalised crosscorrelation coefficient $p_{x,y}(n)$ defined in Equation 3.10. This coefficient is determined for values of n in steps of a decimation factor L (a positive integer excluding zero).

$$p_{x,y}(n) = \frac{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL).y(jL)}{\sqrt{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL)^2 . \sum_{j=1}^{\lfloor n/L \rfloor} y(jL)^2}} \tag{3.10}$$

$$= r_{x,y}(n) \qquad \text{if } L = 1 \tag{3.11}$$

$$\{n = N_{min} + i.L \mid i \in 0, 1, \ldots; N_{min} \leq n \leq N_{max}\}$$

The decimation factor L is used to reduce the computational load of the algorithm. If L is set too low, then calculation of the normalised crosscorrelation coefficient will be computationally expensive and time consuming. If L is set too high, then candidate values for the fundamental period of the frame will be determined with less accuracy. As a compromise, L is set to four for this investigation (equivalent to down-sampling the speech waveform to 5kHz).

Values of $p_{x,y}(n)$ are only valid if there are four or more zero-crossings in the sample set $s_{2n} = \{s(i) \mid i \in 1, \ldots, 2n\}$. This ensures that at least two oscillations occur within the section of data s_{2n}. If there are less than two oscillations in the data, such as for low frequency nasal segments, then the value of $p_{x,y}(n)$ will be high for low values of n. This would erroneously promote fundamental period candidates which correspond to high frequencies.

Candidate values of the fundamental period are obtained by locating peaks in the normalised crosscorrelation coefficient for which the value of $p_{x,y}(n)$ exceeds a specified threshold, T_{srfd}. The threshold is adaptive and is dependent upon the voicing classification of the previous frame and three empirically determined values. If the previously analysed frame is classified as 'unvoiced' or as 'silence' (which is the initial state) then the threshold is set to 0.88. Otherwise, the previous frame must have been classified as 'voiced' and the threshold is set equal to 0.85 times the value of $p'_{x,y}(n'_0)$; where n'_0 is the

fundamental period estimate (in number of samples) of the previous frame, and $p'_{x,y}(n)$ is the set of normalised crosscorrelation coefficients calculated for the previous frame. This product is not permitted to drop below 0.75.

$$T_{srfd} = \begin{cases} 0.88 & \text{if previous frame 'unvoiced' or 'silent'} \\ \max(0.75, 0.85p'_{x,y}(n'_0)) & \text{if previous frame 'voiced'} \end{cases} \tag{3.12}$$

If no candidates for the fundamental period are found, the frame is classified as 'unvoiced'. Otherwise, the frame is assumed to contain 'voiced' speech. In order to find the optimal fundamental period from the set of candidate values, the candidates are listed in order of increasing fundamental period. The candidate at the end of this list represents a fundamental period of n_M, and the m'th candidate represents a period n_m. A second coefficient, $q(n_m)$, is calculated for each candidate. $q(n_m)$ is the normalised crosscorrelation coefficient between two sections of length n_M spaced n_m apart.

$$q(n_m) = \frac{\sum_{j=1}^{n_M} s(j).s(j+n_M+n_m)}{\sqrt{\sum_{j=1}^{n_M} s(j)^2 . \sum_{j=1}^{n_M} s(j+n_M+n_m)^2}} \tag{3.13}$$

The first coefficient $q(n_1)$ is then assumed to be the optimal value. If a subsequent $q(n_m)$ exceeds this optimal value when multiplied by 0.77 (an empirically determined value) then it is in turn assumed to be the optimal value. The candidate for which $q(n_m)$ is believed to be the optimal value forms the estimate for the fundamental period, n_0, of the frame being analysed.

Finally the algorithm obtains an estimate of the fundamental period with a fine resolution. A more accurate fundamental period estimate for the frame is determined by calculating $r_{x,y}(n)$ (Equation 3.9) for n in the region $n_0 - L$ to $n_0 + L$. The location of the maximum within this range corresponds to a more accurate value of the fundamental period. This final estimate is then refined to eliminate the effect of time quantisation errors, by using an interpolation method which is described in Medan *et al.* (1991).

3.2.7 FDA summary

Methods of determining the fundamental frequency of a speech waveform operate either on a frequency domain representation, a time domain representation or on a function of correlation. Time domain and correlation based algorithms make use of filtering techniques in an attempt to enhance performance, whereas the frequency based algorithms use the raw speech waveform. The cepstrum-based (CFD) and harmonic product spectrum (HPS) algorithms operate on a frequency domain representation and differ considerably in the way they estimate the fundamental frequency of a signal from its spectral characteristics. The feature-based (FBFT) and parallel processing (PP) methods operate on a time domain representation. These two methods take a number of measurements directly from the temporal structure of the sampled speech waveform. This has the advantage of being less computationally expensive than the frequency domain based algorithms. The integrated *FØ* tracking algorithm (IFTA) and the super resolution *FØ* determinator (SRFD) use a normalised crosscorrelation function. Although the underlying function is similar for these two algorithms, the way in which they determine fundamental period estimates from the correlation function differs considerably.

The most reliable and accurate method of determining the fundamental frequency of a speech waveform is sought in order to minimise the number of errors, which occur during *FØ* extraction, propagating into the prosodic analysis. A detailed performance evaluation of *FØ* determination algorithms is described in Section 7.2.

3.3 Syllabification

The concept of a syllable revolves round the perceptual prominence of speech sounds in sequence (ie. prominence at a segmental level). The number of syllables in a sequence of sounds is judged by a listener on the basis of the number of peaks of prominence perceived. In general, a peak of prominence is perceived where a speech sound carries greater sonority relative to its nearest neighbours. The production of vowels usually involves less constriction of the vocal tract than in the production of consonants. Vowels, therefore, carry greater sonority and constitute the nucleus of the syllables in which they occur. However, relatively prominent consonants can also take on the nucleus of a syllable

containing no vowels. The behaviour of the speech sounds between syllable nuclei is of no consequence to the number of perceived prominences. The boundaries between syllables are therefore ill-defined.

Each syllable in an utterance holds the potential of being prominent at a suprasegmental level. Hence, the automatic identification of syllables in connected speech forms a part of prosodic analysis. Several schemes have been devised to partition the speech signal into syllable-sized units.

Mermelstein (1975) uses a *loudness function* as a measure of sonority across phones. The loudness function is a time-smoothed and frequency-weighted (-12dB/octave outside the range 500Hz–4kHz) summation of a speech signal's short-time power spectrum. A "convex-hull" algorithm (*ibid.*) is employed to locate dips in the loudness function, which form potential syllable unit boundaries. The algorithm is dependent upon a *significance* parameter which eliminates small dips (below 2dB) in the loudness function as potential boundaries, and a minimum syllable duration of 80ms is imposed. A 90.5% syllable detection rate is reported (6.9% syllables missed and 2.6% extra syllables).

Lea (1980) locates syllable nuclei by seeking dips (4 or 5dB) in a low-band energy contour (summation of energy in the frequency band 60Hz–3kHz). The energy dips are assumed to be associated with pre-vocalic and post-vocalic consonants. The high energy region between dips is considered to be a syllable nucleus if it consists of at least 30ms of voiced speech. The algorithm is reported to identify 90% of syllable nuclei (with 1% extra and 9% missing).

The automatic syllabification of isolated words is conducted by Aull & Zue (1985). Speech is initially segmented into broad phonetic classes — sonorants (vowels, nasals, liquids and glides), unvoiced obstruents, voiced obstruents, voice bars and silence. Segments classified as sonorant are further processed to locate intervocalic nasals, /l/ and /r/, and vowel-vowel transitions. The division of sonorant segments uses spectral weighting functions across selected frequencies which are aimed at capturing formant movements. All undivided sonorant segments and each vocalic region of the divided sonorant segments are taken as syllable nuclei. A performance evaluation of this syllabification scheme is not reported.

Waibel (1988) defines the syllable boundary, for his purposes, as the onset of the

vocalic nucleus and proposes two algorithms to determine such boundaries. He reports a rule-based algorithm which yields a 90–96% syllable detection and a "Zapdash" based algorithm which achieves 90–93% syllable detection. Waibel states, however, that these are results of an informal experiment and gives no indication of the criteria used to syllabify the test data by hand.

The syllabification algorithms cited above are used in speech recognition applications to aid lexical access. Reliable information about the segmental content of an utterance is not available to the syllabification algorithms. In fact, additional information such as syllable structure is needed in speech recognition applications to aid lexical access because reliable information about the segmental content of an utterance is not available. In the application of prosodic analysis for computer aided pronunciation teaching, however, the orthographic transcript of an utterance is known because a foreign language learner is asked to read a given sentence in the course material. Reliable segmental information is therefore available. A syllabification algorithm is described in Section 8.1.1 which uses phone boundary and label information in conjunction with a low-band energy contour.

3.4 The composite structure of *FØ* contours

The raw fundamental frequency contour of a speech waveform does not form an acoustic-phonetic representation of the utterance intonation. It is assumed that a raw *FØ* contour constitutes four superimposed components.

- The intonation pattern of a sentence uttered by different speakers produces *FØ* contours with different fundamental frequency ranges despite the fact that the underlying intonation pattern is identical. Thus, an *FØ* trajectory in different fundamental frequency ranges from different speakers can correspond to the same phonological pitch accent. This speaker-dependent component of an *FØ* contour is a consequence of the anatomical and physiological differences of talkers.

- Macroprosodic variations in an *FØ* contour reflect a speaker's choice of intonation pattern for an utterance. The automatic prosodic analysis of speech aims to isolate this component in order to locate and identify pitch accents.

- Microprosodic variations are imposed on an $F\emptyset$ contour by the segmental content of an utterance (Silverman, 1987). Silverman argues that microprosodic variations of fundamental frequency[4] occur during and in the immediate vicinity of phonetic segments, and that these variations may be greater than macroprosodic $F\emptyset$ variations.

- An $F\emptyset$ contour (produced by any fundamental frequency determination algorithm) can be expected to contain values which are inaccurate. An $F\emptyset$ contour also consists of cycle-to-cycle jitter ($F\emptyset$ perturbations) generated at the vocal folds (Hiller, 1985). Methods to reduce errors resulting from the automatic determination of fundamental frequency and the smoothing of $F\emptyset$ perturbations are investigated in Section 7.3.

Two schemes to compensate for between-speaker differences in fundamental frequency range are reported as relatively successful in a review of normalisation methodologies by Rose (1987).

The first scheme, fraction of range normalisation, involves expressing an observed $F\emptyset$ value as a fraction of the difference between range-defining $F\emptyset$ values,

$$F\emptyset_{norm} = \frac{F\emptyset_{input} - F\emptyset_{min}}{F\emptyset_{max} - F\emptyset_{min}} \tag{3.14}$$

This scheme has the disadvantage that calculating satisfactory range-defining parameters (that are assumed to be equivalent between speakers) is difficult without first completing the normalisation.

In the second scheme, z-transform normalisation, an observed $F\emptyset$ value is expressed as a multiple of a measure of dispersion relative to the mean fundamental frequency. The normalised $F\emptyset$ value is given by,

$$F\emptyset_{norm} = \frac{F\emptyset_{input} - \overline{F\emptyset}}{\sigma_{F\emptyset}} \tag{3.15}$$

[4]Silverman (1987) refers to microprosody as *segmental perturbations* in order to emphasis that this component of an $F\emptyset$ contour can influence the perceived naturalness of synthesised speech. He argues that the term *microprosody* misleadingly implies that this component is smaller than suprasegmental (macroprosodic) variations. Whilst this thesis acknowledges the importance of microprosody in the perceived naturalness of synthesised speech, the term segmental perturbations is not used in order to avoid confusion with $F\emptyset$ perturbations related to cycle-to-cycle jitter.

where $\overline{F\emptyset}$ is the long term mean fundamental frequency for a given speaker, and $\sigma_{F\emptyset}$ is the long term population standard deviation. Thus the normalised $F\emptyset$ contour is centred around the zero with each unit representing a frequency which is one population standard deviation from the mean. This scheme has the advantage of reflecting a relatively stable statistical distribution from a large number of fundamental frequency values, rather than just the two range-defining parameters as is the case with fraction of range normalisation.

The *stylisation* of an $F\emptyset$ contour aims to isolate the macroprosodic component by removing the microprosodic component for a given speaker, under the assumption that no $F\emptyset$ errors exist because of malfunctions in an FDA. A requirement of the $F\emptyset$ stylisation is to ensure that any microprosodic $F\emptyset$ variations which are larger than macroprosodic $F\emptyset$ variations are prevented from being confused as pitch accents, and that any macroprosodic $F\emptyset$ variations which are smaller than microprosodic $F\emptyset$ variations are not removed.

A number of algorithms have been proposed to automatically stylise an $F\emptyset$ contour as a sequence of piece-wise straight lines (Scheffers, 1988), quadratic spline curves (Hirst & Espesser, 1993) or rise-fall-connection elements (Taylor & Isard, 1992; Taylor, 1993) where the rise elements and fall elements are modelled by a polynomial function, and the connection element is a straight line. $F\emptyset$ contours which are stylised as a sequence of straight lines by hand are used in perceptual studies of the intonation of several European languages, including Dutch (Cohen & 't Hart, 1967; 't Hart & Cohen, 1973), British English (de Pijper, 1983) and French (Beaugendre *et al.*, 1992). It is argued in these studies that a sequence of straight lines produces a "close-copy" stylisation such that speech resynthesised using the stylised $F\emptyset$ contours cannot be distinguished perceptually from speech resynthesised using the original $F\emptyset$ contours. A comparative investigation into the use of straight line and parabola stylisation is reported by 't Hart (1991).

In the generation of an acoustic-phonetic representation of prosodic aspects of speech for computer aided pronunciation teaching, the stylisation of an $F\emptyset$ contour aims to remove the microprosodic component of the contour. It is therefore immaterial as to whether or not the stylisation is based on straight lines or on polynominal functions. Linear piece-wise stylisation is employed in Section 8.2.1.

3.5 Automatic transcription of prosodic structure

Algorithms to extract prosodic features from the speech waveform have previously been developed for speech recognition applications. The motivation behind the development of such algorithms has been the notion of islands of reliability (Lea, 1980; Aull & Zue, 1985) where phonetic information is assumed to be more robust. The locations of such islands of reliability are provided by prominent syllables. Stress information is also required to aid lexical access. A review of the use of prosodic structure in automatic speech recognition systems is presented by Vaissière (1988).

The need to automatically extract prosodic features from a speech waveform is motivated by requirements other than speech recognition and computer aided pronunciation teaching applications. It is also potentially valuable for linguistic studies of corpora which are too large to analyse consistently by hand. The automatic transcription of prosodic aspects of speech, if designed on a cogent linguistic basis, is a useful tool for developing a better understanding of the relation between prosodic and syntactic structures and their phonetic realisation.

Lea (1974; 1980) describes an algorithm to annotate syllables as either prominent or non-prominent. The algorithm partitions an utterance into *syntactic constituents* on the basis of fluctuations in its *FØ* contour. Constituent boundaries are placed both at the minimum point in a valley of the *FØ* contour for which there is at least a 7% change in *FØ* at both sides, and at periods of unvoiced speech greater than 350ms. The *FØ* contour within each constituent is assumed to initially rise to some maximal point, then gradually fall. A straight *threshold* line (on a logarithmic scale) is taken from the initial peak in *FØ* to the end of the final fall in the constituent. Automatically located syllable nuclei (see Section 3.3) are annotated as prominent if they are associated with the initial *FØ* rise of a constituent, or if the local *FØ* rises above the threshold line. This algorithm is dependent on the performance of the *FØ* determination algorithm and is prone to error because of the unreliable automatically identified location of constituent boundaries. Lea's algorithm is reported to annotate 85% of syllables as prominent or non-prominent in agreement with human perception.

The algorithm proposed by Lea does not use duration, energy or vowel quality to locate prominent syllables. It terms of the definition of prominence outlined in Sec-

tion 2.2.4, this algorithm locates pitch accents which are associated with prominent syllables, but it does not locate prominent syllables *per se.*

Aull & Zue (1985) propose an algorithm to determine lexical stress from the acoustic waveform for multi-syllabic isolated words. Syllables are annotated as prominent, non-prominent or reduced with a reported 87% accuracy. A vector of five normalised features is calculated for each syllable. The duration of the syllable nucleus (adjusted for pre-pausal lengthening according to phonetic context); the logarithmic average syllable energy in the frequency bands 400Hz–5.0kHz and 1.2–3.3kHz; the peak fundamental frequency; and a measure of spectral stability are used as features. Syllables bearing stress are believed to have spectrally stable nuclei (except diphthongs). A reference vector is formed from the maximum value of each vector across the syllables of a word. The syllable with the minimum Euclidean distance from its feature vector to the reference vector is designated as prominent. The distinction between non-prominent and reduced syllables is made using the duration and energy features alone. The Euclidean distance provides a measure of the *degree of prominence* of a syllable relative to other syllables in a word. This can be used to provide second candidates for the prominent syllables in multi-syllabic words (secondary stress).

The use of a Bayesian classifier (Appendix B) to assign a *probability of stressedness* to syllables in connected speech, rather than a binary decision (prominent or non-prominent) is described by Waibel (1988). In evaluating the classifier, however, any syllable with a higher probability of being prominent than non-prominent is assigned as prominent, otherwise it is assigned as non-prominent. A vector of acoustic features is used as the input to the classifier. There are a number of possible acoustic features which can be selected to form the input vector for each syllable.

- Integral of the peak-to-peak amplitude over the vocalic portion of the syllable (*ptpint*).
- Duration of the vocalic portion of the syllable (*sondur*).
- Duration of the entire syllable from the onset of the initial vowel to the beginning of the next syllable (*syldur1*).
- Duration from the end of the initial vowel of the previous syllable to the end of

the initial vowel in the syllable concerned (*syldur2*).

- Maximum *FØ* during the syllable (*F0max*).
- Average *FØ* during the syllable (*F0ave*).
- Offset in *FØ* from the syllable concerned to the following syllable (*F0offs*).
- Average spectral change (Aull & Zue, 1985) over the syllable nucleus (*spchave*).

Waibel uses a training database of connected speech consisting of syllables assigned as either prominent (primary and secondary stress) or non-prominent from their lexical assignment to investigate the correlation of the acoustic features with lexical stress. This approach differs from the need for prosodic analysis in computer aided pronunciation teaching proposed in Chapter 2 in that not all lexical stress placements appear as sentential stress in phrases of connected speech. Waibel clearly stipulates that the stress detection algorithm he proposes, "ignores sentential stress, emphasis, phrase level phenomena, and rhythmic/syntactic/semantic phenomena" (*ibid*, pp.108).

Bayesian classification is initially performed with an input vector containing just one of the acoustic features. Waibel's experimental results show that for this unidimensional classifier, *sondur* and *F0max* generate the least number of classification errors of all the duration and fundamental frequency features, respectively. However, in Waibel's experiments using a Bayesian classifier with a multi-feature vector, he does not report the combination of *ptpint*, *sondur*, *F0max* and *spchave* as a feature vector. A recalculation of Waibel's performance evaluation was conducted, as part of this thesis, to determine the performance of the Bayesian classifier under an open test (by excluding the database used in training the Bayesian classifier). The recalculation shows the Bayesian classifier, at best, to achieve 86.6% correct classification by using the features, *ptpint*, *syldur1*, *F0max* and *spchave*. The use of any additional features reduces the percentage of correct classification. The next best performance of 86.4% is achieved by using the features *ptpint*, *syldur1*, and *spchave*. This is contrary to Waibel's conclusion which is based on an evaluation which includes the training database (closed test). There is no evidence to support that the difference in performance of just five syllables (out of 2368) is statistically significant. Therefore, the introduction of the *FØ* feature has little effect on the

classifier's performance. This is surprising given that so much literature (Section 3.1) has indicated that fundamental frequency and the placement of pitch accents are correlated with stress. It is possible that the features *F0max*, *F0ave* and *F0offs* do not capture the particular aspects of an *FØ* contour which are associated with pitch accents and subsequently with the location of prominent syllables.

The features that are selected to give the above 86.6% correct classification rate describe only the acoustic parameters for the syllable in question. The variation of these parameters relative to neighbouring syllables is not included as an input parameter to the classifier. This exclusion cannot be beneficial to the classifier, given that a syllable's prominence is relative to its neighbours by definition. (The prominence of syllable in isolation has no meaning.) In a system designed to extract prosodic information in French, Vaissière (1989) uses acoustic parameters which also describe parameter variations from one syllable to the next.

A move towards linguistic knowledge-based rules to locate prominent syllables is made by Hieronymus (1989; 1991). The acoustic parameters, duration, energy, and fundamental frequency are processed in parallel to determine their unidimensional contribution to the perceived stress of each vowel in an utterance. A measure of vowel quality or spectral stability is not used and a syllabification stage is not included. It is assumed that every vowel in an utterance forms the nucleus of a syllable and that no syllabic consonants form the nucleus of a syllable. The analysis system proposed by Hieronymus manipulates the duration, energy and *FØ* acoustic parameters in the following way.

Duration: The duration of the final vowel preceding either a section of speech labelled as a pause or the end of an utterance, is reduced by a fixed factor (0.6) to compensate for pre-pausal lengthening. Each vowel is classed on a broad phonetic basis as either a short vowel, a long vowel or a diphthong, and its duration is regarded as contributing to the prominence of vowel if it is greater than an *a priori* threshold duration respective of vowel-type (short, long or diphthong). A limit is imposed on the percentage of the vowels in the utterance which are marked as potentially prominent on account of duration.

Energy: The maximum low-band energy in each vowel is used as a measure of its intensity. The average of the two highest vowel intensities in the utterance is used as a

reference. A vowel with an intensity within 7dB of the reference is regarded as having sufficient energy to contribute to its perceived prominence. Vowels with an intensity more than 20dB from the reference are taken to be definitely non-prominent, regardless of the other acoustic parameters.

Fundamental frequency: The *FØ* contour is divided into sections, one section for each vowel. The *FØ* section associated with a vowel V_0 is a region of continuously voiced speech which runs from the end of the first unvoiced consonant preceding the vowel to the beginning of the first unvoiced consonant succeeding the vowel. If another vowel V_{-1} is encounted between the preceding unvoiced consonant and vowel V_0, then the *FØ* section starts from the mid-point between the end of vowel V_{-1} and the beginning of vowel V_0. Similarly, if another vowel V_{+1} is encounted between the succeeding unvoiced consonant and vowel V_0, then the *FØ* section ends at the mid-point between the end of vowel V_0 and the beginning of vowel V_{+1}.

The peak *FØ* value and the *FØ* values at the beginning and at the end of each section are determined. The initial and final two *FØ* estimates are ignored during this process because they can be highly influenced by the potentially unstable behaviour of the vocal folds at the onset and offset of voicing. The overall slope is calculated from the locations and values of the *FØ* readings at the beginning and the end of each section. Each *FØ* section is categorised as a 'fall', 'rise', or 'level' on the basis of the gradient of this slope.

$$\text{Slope-type} = \begin{cases} \text{'fall'} & \text{if } \textit{gradient} < -100\text{Hz/second} \\ \text{'rise'} & \text{if } \textit{gradient} > 100\text{Hz/second} \\ \text{'level'} & \text{otherwise} \end{cases} \tag{3.16}$$

Each vowel is classified as either 'pitch accented' ("*PA*") or 'unaccented' ("*UA*") using the decision filter illustrated in Figure 3.5. The decision filter examines three consecutive *FØ* sections for each vowel — the *FØ* sections associated with the vowel and its left and right contexts. The first and last *FØ* section of an utterance do not have a left and a right context respectively. These end points are represented by 'NA' (not available). Figure 3.5 shows all three-way combinations of fall, rise, level and 'NA' which can arise and the corresponding output of the decision filter.

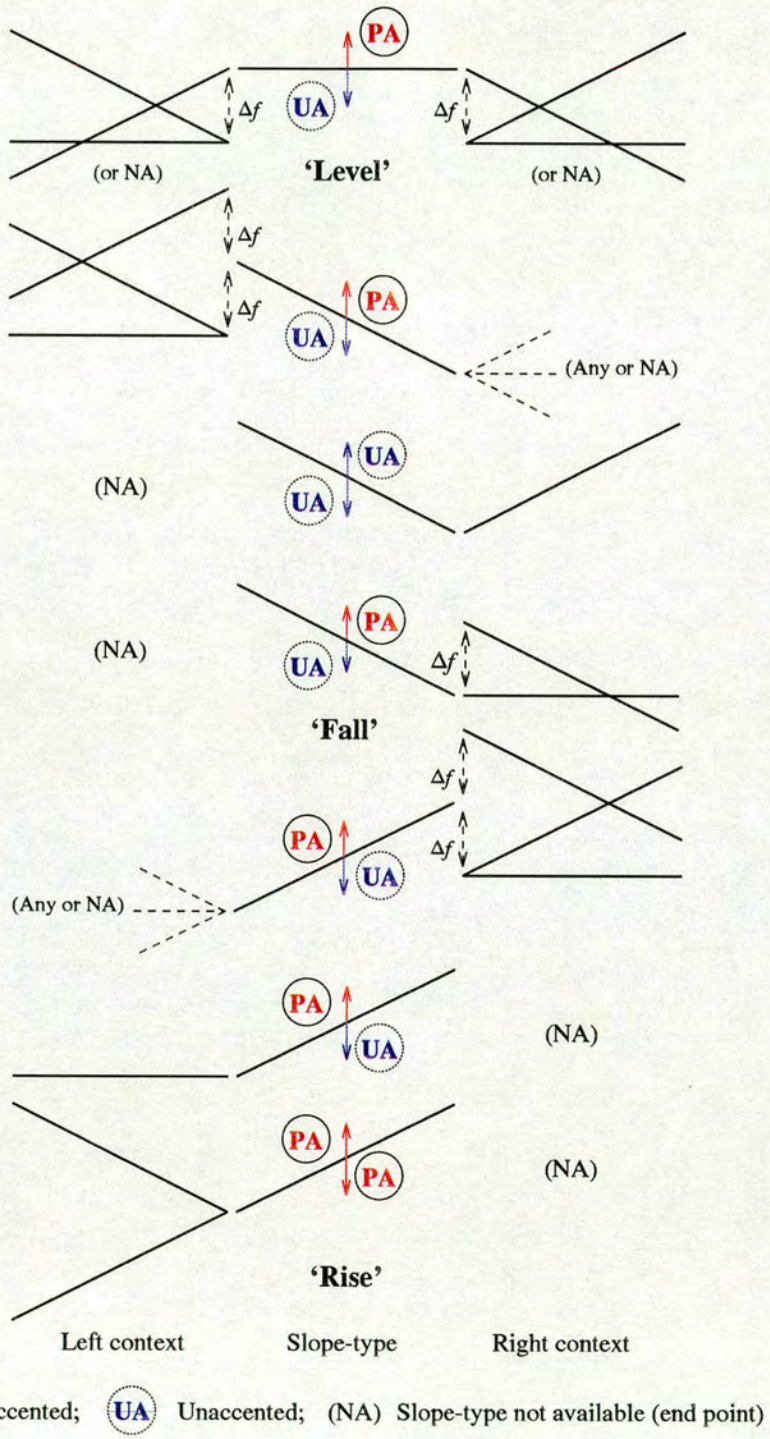


Figure 3.5: Pitch accent decision filter

A vowel associated with a level *FØ* section is classified as pitch accented if, for example, its left context is either falling or level and there is a step up of at least Δf (Hz)[5] from the end of the left context to the beginning of the level *FØ* section, and if the right context is a fall and the end of the level *FØ* section is not less than the beginning of the right context. If, however, the step up from the end of the falling or level left context to the beginning of the level *FØ* section is less than Δf (Hz) or if there is a step up from the end of the *FØ* section to the beginning of the right context, then the vowel is classified as unaccented.

Similarly, a vowel associated with a falling *FØ* section is classified as pitch accented if, for example, its left context is a rise and the beginning of the falling *FØ* section is no more than Δf (Hz) less than the end of the rising left context (the beginning of the falling *FØ* section can be any amount higher than the end of the rising left context). The right context is ignored in cases of a falling *FØ* section. If, however, the step down from the end of the rising left context to the beginning of the falling *FØ* section is more than Δf (Hz) then the vowel is classified as unaccented.

If a vowel is associated with a falling *FØ* section and the left context is not available (because the first vowel of an utterance is being classified) then the vowel is always classified as unaccented when it is followed by a rise.

The classification of a vowel as either pitch accented or unaccented can be determined from Figure 3.5 for any combination of fall, rise, level and 'NA' in the *FØ* sections associated with a vowel and its left and right contexts.

The pitch accent decision filter developed by Hieronymus has a number of advantages over the *FØ* features used in the Bayesian classifier technique. Firstly, judgements of vowel accentuation are based on trajectories of *FØ* in and around the vowel, rather than being based on static, isolated *FØ* measurements. Secondly, listeners classify syllables as accented if a pitch discontinuity is perceived. The pitch accent decision filter is designed to capture such discontinuities. This technique of locating pitch accents is therefore used in the system described in Section 8.2.2.

Prominence judgements for vowels are based on combining the unidimensional contribution to the perceived stress of each vowel obtained from the analyses of duration,

[5]In the implementation of the pitch accent decision filter, Δf is set to 9Hz.

energy and fundamental frequency outlined above. A vowel in an utterance is categorised as prominent if at least two-out-of-three of the acoustic parameters are indicative of stress in an equal-weight voting procedure. Only the energy parameter is given the power to exercise veto against a combined vote for prominence by the duration parameter and an *FØ* 'pitch accented' label. Hieronymus (1991) reports that this procedure agrees with 77.2% of human judgements of vowel prominence in connected speech.

Wightman & Ostendorf (1991; 1992) focus on the automatic recognition of prosodic boundaries and intonational features as well as determining the relative prominence of syllables. Linguistically motivated features are used to automatically determine the break index between adjacent words and to label each syllable as one of four categories — 'default', pitch accented syllable, phrase-final syllable with boundary tone, and boundary tone also perceived as pitch accented. The features used are:

Word dependent information: The lexical stress of a syllable; and a flag indicating the location of word-final syllables. The phonetic transcription of an utterance is determined by an automatic segmentation algorithm given an orthographic transcription of the sentence. This provides word boundary information which is combined with a lexicon to determine the lexical stress of syllables. This algorithm therefore uses the information which Waibel's algorithm tries to determine, as one of its input features.

Duration: The duration of a pause (if one exists) at the end of a word; the average z_{mean} duration measure of phones in the rhyme of the final syllable in a word; and the difference in the average z_{mean} duration measure for the syllable rhyme and the syllable onset (see Chapter 4 for details of syllable structure and z-score duration measures).

Energy: The mean energy in a syllable. The energy at the end of an utterance is often observed to reduce simply because the speaker is 'running out of breath'. It therefore has the potential to being used as a cue for the end of a phrase.

Fundamental frequency: The initial, final, maximum, minimum and average *FØ* val-

ues are determined for each syllable i. The following relative values are calculated:

$$F\emptyset_{max}(i)/F\emptyset_{average}(i+1)$$

$$F\emptyset_{max}(i)/F\emptyset_{max}(i-1)$$

$$F\emptyset_{max}(i)/F\emptyset_{average}(i)$$

$$F\emptyset_{min}(i)/F\emptyset_{average}(i)$$

$$F\emptyset_{final}(i)/F\emptyset_{average}(sentence)$$

The shape of the $F\emptyset$ contour during the course of a syllable is described as either a rise, a fall, a rise-fall or a fall-rise by comparing the initial, average (mid) and final $F\emptyset$ values.

These features form the input to a seven-state (one state for each level of break index) continuous Hidden Markov Model (HMM) and to a four state (one state for each syllable category) discrete HMM. The HMMs are fully connected in the absence of a theory either on the interaction between break indices or on the sequences of syllable categories. In evaluating this technique, Wightman & Ostendorf report that 77% of boundary tone are correctly detected with a 3% false detection rate, and that 86% of pitch accents are correctly detected with a 14% false detection rate.

3.6 Summary

The acoustic parameters which are identified by researchers as correlates of prosodic phenomena are duration, energy, fundamental frequency, and vowel quality. There is relatively little consideration of the effects of vowel quality on the perception of (lexical or sentential) stress compared with the extensive research relating duration, energy and $F\emptyset$ with stress.

Studies of the acoustic correlates of stress are not consistent in the features they measure. 'Duration' is used by different researchers to refer to the duration of the vocalic portion of a syllable (Fry, 1955; Fry, 1958; Waibel, 1988) and to the duration of an entire syllable (Lieberman, 1960; Adams & Munro, 1978) with little consideration to the definition of a syllable (neither a phonetic definition nor a phonological definition). 'Energy' is used to refer to the peak amplitude in a syllable (Fry, 1955; Lieberman,

1960), to the amount of fall from a peak in the amplitude envelope of a syllable (Adams & Munro, 1978), to the integral of the peak-to-peak amplitude over the vocalic portion of a syllable (Waibel, 1988), to the average low-band syllable energy (Aull & Zue, 1985), and to the maximum low-band vowel energy (Hieronymus, 1989; Hieronymus & Williams, 1991). 'Fundamental frequency' is used as an acoustic correlate of stress. However, fundamental frequency is principally related to the characteristic melody of an utterance. Measures of *FØ* in a syllable are used as acoustic correlates of stress with little or no consideration of the intonational role of fundamental frequency. This approach is only viable in identifying the prominent syllables of isolated words. Fundamental frequency can be used as a secondary cue to the location of prominent syllables in connected speech because pitch accents are observed to fall on prominent syllables.

The fundamental frequency of a speech waveform must be determined as an initial process for prosodic analysis. The functionalities of a selection of *FØ* determination algorithms are described in detail. The performance of FDAs must be considered with respect to their application in systems of prosodic analysis and errors arising from the malfunctions of FDAs must be prevented from propagating into the subsequent prosodic analysis of speech.

Many of the duration and energy features are related to the definition of a syllable. Furthermore, prosodic aspects of speech are described in a syllabic domain. The automatic syllabification of speech is therefore an important part of prosodic analysis. Algorithms devised to partition a speech signal into syllable-sized units use measures of sonority based on low-band energy and use spectral characteristics to locate boundaries between adjacent sonorants. The reviewed algorithms are applied to speech recognition systems. Reliable information about the segmental content of an utterance is therefore not available to them. In the application of prosodic analysis for computer aided pronunciation teaching, however, reliable segmental information is available and may be used to enhance the syllabification process.

An emphasis is placed on the fact that the acoustic parameters which are associated with prosodic aspects of speech are also influenced by non-prosodic aspects of speech and that they interact in the manifestation of stress. It is highlighted that the *FØ* contour of an utterance is affected by the talker's anatomy and physiology (speaker-dependent

FØ range), the segmental content of an utterance (microprosodic variations), and cycle-to-cycle jitter (*FØ* perturbation) together with errors involved in its determination from the speech waveform.

The reviewed algorithms for prosodic analysis have not addressed the problems of determining the fundamental frequency of speech and have not comprehensively addressed the need to normalise for non-prosodic variations in the acoustic parameters which are used in the prosodic analysis. The system of prosodic analysis described in this thesis addresses the problems related to the determination of the fundamental frequency of speech and focuses on techniques of normalising for variations in acoustic parameters which are due to non-prosodic aspects of speech.

Chapter 4

Duration measures

The acoustic parameters used in the analysis of prosodic aspects of speech must be extracted from a speech waveform. There are four acoustic parameters which are correlated with prosodic phenomena — duration, energy, vowel quality and fundamental frequency. This Chapter and Chapters 5, 6 & 7 concentrate on the extraction of each of these acoustic parameters from a speech signal.

The domain of phonetic units whose duration are to be determined as optimal correlates of stress and normalisation techniques applied to them are investigated in this Chapter. The phonetic units investigated are related to syllable structure. The normalisation techniques aim to compensate for variations in duration which are due to non-prosodic aspects of speech. The underlying principle of these investigations is to normalise the acoustic parameters for non-prosodic aspects of speech such that the processed acoustic parameters can then be combined to form a prosodic description of speech.

4.1 Database description

Two databases of hand labelled phonetically balanced sentences are used in the experimental investigations of duration measures (presented in this Chapter) and of energy measures and vowel quality measures (presented in Chapters 5 & 6 respectively). The first database (referred to as the *training data*) consists of 200 sentences read by a male speaker of British English (South-eastern dialect) with a non-pathological voice. The read sentences are recorded in an anechoic studio with a close-talking microphone and a

16-bit analogue-to-digital converter sampling at 20kHz. The second database (referred to as the *test data*) consists of 460 sentences read by the same speaker under the same conditions. The speaker was unaware during recording that the utterances would be used in a study of prosodic analysis.

The training data sentences are a set of 200 'phonetically rich' sentences designed so as to provide almost total coverage of permissible demi-syllables in English (Laver *et al.*, 1988). The test data sentences are a set of 460 'phonetically compact' sentences designed so as to provide as complete a coverage of phoneme pairs as possible (Lamel *et al.*, 1986). These 460 sentences are the same as the anglicised TIMIT sentences used in SCRIBE (Spoken Corpus Recordings In British English). All the speech data is phonetically transcribed by phoneticians using the criteria proposed by Laver *et al.* (1989).

Phones are grouped into syllabic units automatically using the algorithm described in Section 8.1 and the perceived prominence of each syllable is transcribed by hand. The number of degrees of stress that constitute an adequate description of speech is disputed amongst linguists (Section 2.2.1). The practical approach taken here is to describe syllables as either prominent or non-prominent. If they are prominent, they may or they may not be associated with pitch accents. Pitch accented prominent syllables are further classed into nuclear and non-nuclear accents. Thus, syllables perceived as bearing sentential stress (Section 2.1.1) are transcribed either as nuclear accented "*n*" (primary stress), as non-nuclear accented "*a*" (secondary stress) or as unaccented but stressed "*s*" (tertiary stress); otherwise they are transcribed as unstressed "*u*"[1] (Bagshaw & Williams, 1992). Nuclear and non-nuclear accented syllables are collectively referred to as pitch accented "*PA*" in this thesis. Pitch accented and stressed (but unaccented) syllables are collectively referred to as prominent "*P*". In contrast to prominent syllables, unstressed syllables are also referred to as non-prominent "*NP*".

The transcription of the perceived prominence of each syllable in the training and test data is performed by the author (who is also a native speaker of British English). There are inevitably instances when the discrete categorisation of a syllable as prominent or non-prominent is ambiguous. Hence, the transcription of the data cannot be regarded as definitive. There is therefore some degree of error in the comparisons of prominence

[1] See Section 2.2.4 for a definition of these terms.

levels labelled by hand and labelled by the automatic procedures discussed below. In the prosodic labelling of the Lancaster/IBM spoken English corpus, two phoneticians working independently to transcribe syllables as either prominent "*P*" or non-prominent "*NP*" are reported to achieve 90.8% agreement (Pickering *et al.*, 1994). The transcribers are reported to achieve 83.1% agreement in categorising syllables as either pitch accented "*PA*", stressed "*s*" or unstressed "*u*".

4.2 Phonetic units

The term *duration* refers to the length of a particular constituent of speech. In the context of prosodic analysis, the constituent of speech is usually a type of phonetic or phonological unit. Some level of abstraction from the acoustic waveform has, therefore, already been applied to derive the unit and so duration is not strictly an acoustic parameter. However, duration is widely regarded as a fundamental acoustic correlate of prosodic phenomena (see Section 3.1). The units whose durations best correlate with sentential stress and normalisation techniques which can be applied to them, are investigated here. The aim of the investigation is to determine a duration feature whose distributions for each prominence level have the greatest separability.

The units whose durations are investigated are:

- U_1, the syllable nucleus.
- U_2, syllable nucleus and coda (the syllable *rhyme*).
- U_3, the syllable onset and nucleus (which is referred to hereinafter as the syllable *lhyme* — a 'left-hand' equivalent to the rhyme).
- U_4, the entire syllable (onset, nucleus and coda).
- U_5, a nucleus-to-nucleus unit.

All the phones within a syllable preceding the nucleus form the onset, and all those succeeding the nucleus form the coda. The nucleus-to-nucleus unit starts at the beginning of a syllable nucleus and ends either at the beginning of the nucleus of a following syllable, or, if it is followed by a pause, at the end of the syllable coda. The structures of these

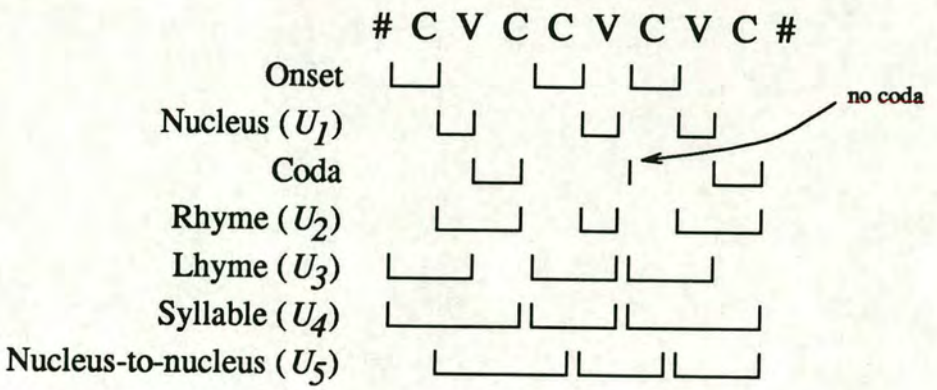


Figure 4.1: Syllable structure

units are illustrated in Figure 4.1. The clustering of phones into syllable-type units is performed automatically using the algorithm described in Section 8.1. The syllable nuclei are also identified by this algorithm.

4.3 Normalisation techniques

The duration of a unit is influenced by many parameters other than just its relative prominence. A normalisation technique is required to compensate for any variations in the duration of a unit that result from such parameters. Several approaches to duration normalisation are investigated.

Klatt (1979) describes a rule based system to estimate phonemic segment durations for synthesised speech in American English. The application of a set of rules modifies the "inherent" durations (specified in a look-up table) of each segment according to its phonemic context, its degree of stress and its proximity to syntactic junctures (word, phrase and clause boundaries). The prosodic analysis of an utterance involves determining the contribution made by stress to the duration of a segment. This may be achieved by employing Klatt's rules to predict the duration of a segment in an unstressed context. The difference between the predicted duration and the actual duration of a segment is a theoretical estimate of the contribution made by stress to the segment duration. The phonetic context of each segment and the syntactic structure of the utterance need to be

known in order to predict the duration of a segment in this way. However, the syntactic structure of an utterance is usually not readily available from the speech waveform (without reliable speech recognition and syntactic parsing). Furthermore, there is no simple way of automatically determining either the factors by which durations are modified under different conditions or the inherent segment durations for a given speaker. This approach is therefore not used in the prosodic analysis system proposed in this thesis.

Fant (1988) proposes the use of duration indices for units in a study of sentential stress in Swedish. A duration index S_i is derived from the measured duration of a unit T_{unit} consisting of i phones, and two reference durations, $\overline{T_u}(i)$ and $\overline{T_s}(i)$. $\overline{T_u}(i)$ is the average duration of non-prominent units which consist of i phones, and $\overline{T_s}(i)$ is the average duration of prominent units which consist of i phones. The reference durations are empirically determined from the training data and are calculated separately for each number of phones i that can exist within a unit.

$$S_i = 1 + \frac{T_{unit} - \overline{T_u}(i)}{\overline{T_s}(i) - \overline{T_u}(i)} \tag{4.1}$$

Campbell & Isard (1991) note that the distributions of segment durations differ by phone-type in terms of both their mean and population standard deviation values. The individual phone durations are normalised by applying a z_{mean}-transform which expresses the phone duration T_{phone} in terms of the mean duration for that phone-type $\mu_T(phone\_type)$ and standard deviation for that phone-type $\sigma_T(phone\_type)$. The values of $\mu_T(phone\_type)$ and $\sigma_T(phone\_type)$ are empirically determined from the training data.

$$z_{mean} = \frac{T_{phone} - \mu_T(phone\_type)}{\sigma_T(phone\_type)} \tag{4.2}$$

The z_{mean} duration measure for a segment will be either positive, if the segment duration is longer than the mean duration for the given phone-type, or negative, if the segment duration is shorter than the mean. Segment durations are thus offset relative to the mean durations of their phone-type and scaled relative to the spread of their distributions. As noted by Wightman *et al.* (1992), a positive z_{mean} duration measure does *not* necessarily mean that a segment is lengthened (due to syllable prominence and/or prosodic phrasing). This is because the mean duration $\mu_T(phone\_type)$ is deter-

mined from all occurrences of a particular phone, including unlengthened ones. There is generally an inequality in the number of lengthened and unlengthened phones. Thus, if there are more lengthened segments than unlengthened segments then there will be a tendency for the mean duration to be somewhat larger than the minimum duration of lengthened phones. Therefore, lengthened segments may have a negative z_{mean} duration measure. Conversely, if there are more unlengthened segments than lengthened segments then there will be a tendency for the mean duration to be somewhat smaller than the maximum duration of unlengthened phones. Therefore, unlengthened segments may have a positive z_{mean} duration measure.

The normalisation procedure which Fant (1988) proposes does not take into account the variance of segment duration as a function of phone-type, and the Campbell & Isard (1991) normalisation method does not take into account the variance of segment duration as a function of the relative prominence of a unit or the number of phones in a unit.

It is proposed in this thesis that the z-transform used by Campbell & Isard can be modified so that it also takes the relative prominence of a unit into account. To do this, first assume that the duration of a phone in a prominent unit is greater than the duration of the same phone-type in a non-prominent unit. Thus, in a distribution of durations for a given phone-type containing p percent non-prominent phones, the lower p percent of durations are assumed to correspond to the durations of the non-prominent phones. Instead of offsetting the segment durations relative to the mean durations of their phone-type, the durations are expressed relative to the p'th-percentile duration for that phone-type $\varphi_T^p(phone\_type)$. This gives rise to a $z_{percentile}$-transform,

$$z_{percentile} = \frac{T_{phone} - \varphi_T^p(phone\_type)}{\sigma_T(phone\_type)} \tag{4.3}$$

If the assumption is correct, then the $z_{percentile}$ duration measure is positive for all prominent phones and negative for all others.

The prominent units in the training database are indicated by a transcription on the nucleus of the unit. It is assumed that the prominence level of a unit only affects the duration of the phones within that unit. Therefore, a non-nuclear phone inherits the prominence level of the nucleus in the same unit for the purpose of calculating its

p'th-percentile. All non-nuclear phones that exist outside the definition of the unit are assumed to be non-prominent.

An underlying assumption of the z_{mean} normalisation (Equation 4.2) is that the populations of duration measurements for phone-types are Normal (Gauss-Laplacian) distributions. These distributions are modelled by the two parameters $\mu_T(phone\_type)$ and $\sigma_T(phone\_type)$. The populations of duration measurements for phones are shown in Figures 4.2 (vowels) & 4.3 (consonants). These populations have positively skewed distributions and can, therefore, only be modelled approximately by the mean and standard deviation parameters. A more robust set of parameters is desired to model these distributions. The $z_{percentile}$ normalisation (Equation 4.3) effectively adjusts the mean parameter by considering the population of duration measurements for a phone-type as constituting two sub-populations — one population of the non-prominent phones and another population of prominent phones.

The duration of each phone, T_{phone} is determined directly from the phonetic transcription. This may be a transcription made either by hand or by some automatic procedure. The z-transformed phone durations (either z_{mean} or $z_{percentile}$ duration measures) derived from these are listed in chronological order to form a duration contour for each utterance. A duration contour can be smoothed using a non-linear smoother (Section 7.3.1) with a window length of 3 phones. Such smoothing aims to iron out boundary placement errors between each pair of phones. Boundary placement errors in auto-segmented data are superimposed on time quantisation errors which can be as much as 5ms. In order to test whether smoothing the duration contour is or is not beneficial to the prosodic analysis, the z-transform normalisation techniques are investigated both without smoothing (S_0) and with smoothing (S_1). Note, however, that the training data and test data used in this investigation are transcribed by hand. Quantisation errors are therefore negligible.

Some phones occur infrequently in the training data. There is therefore the possibility that the sample of phones in the training data is unrepresentative of the universal set of phones. In order to improve the robustness of the data, it may be possible to classify phone-types on a broad phonetic basis rather than on a fine phonetic basis. The duration distributions for phones classified on a fine phonetic basis are shown in Figures 4.2 (vowels) & 4.3 (consonants). Phones may be grouped into broad phonetic classes on the

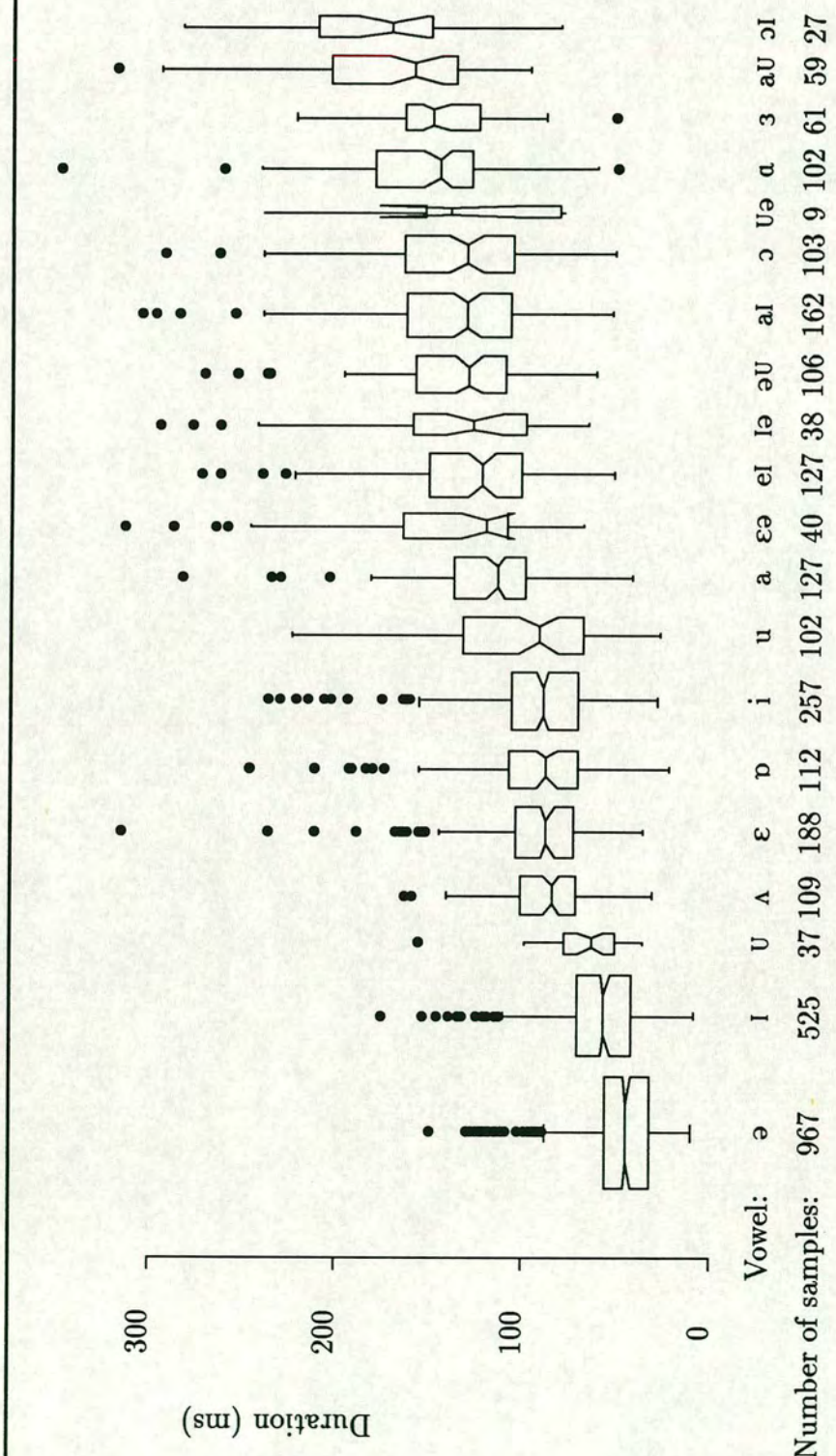


Figure 4.2: Vowel durations

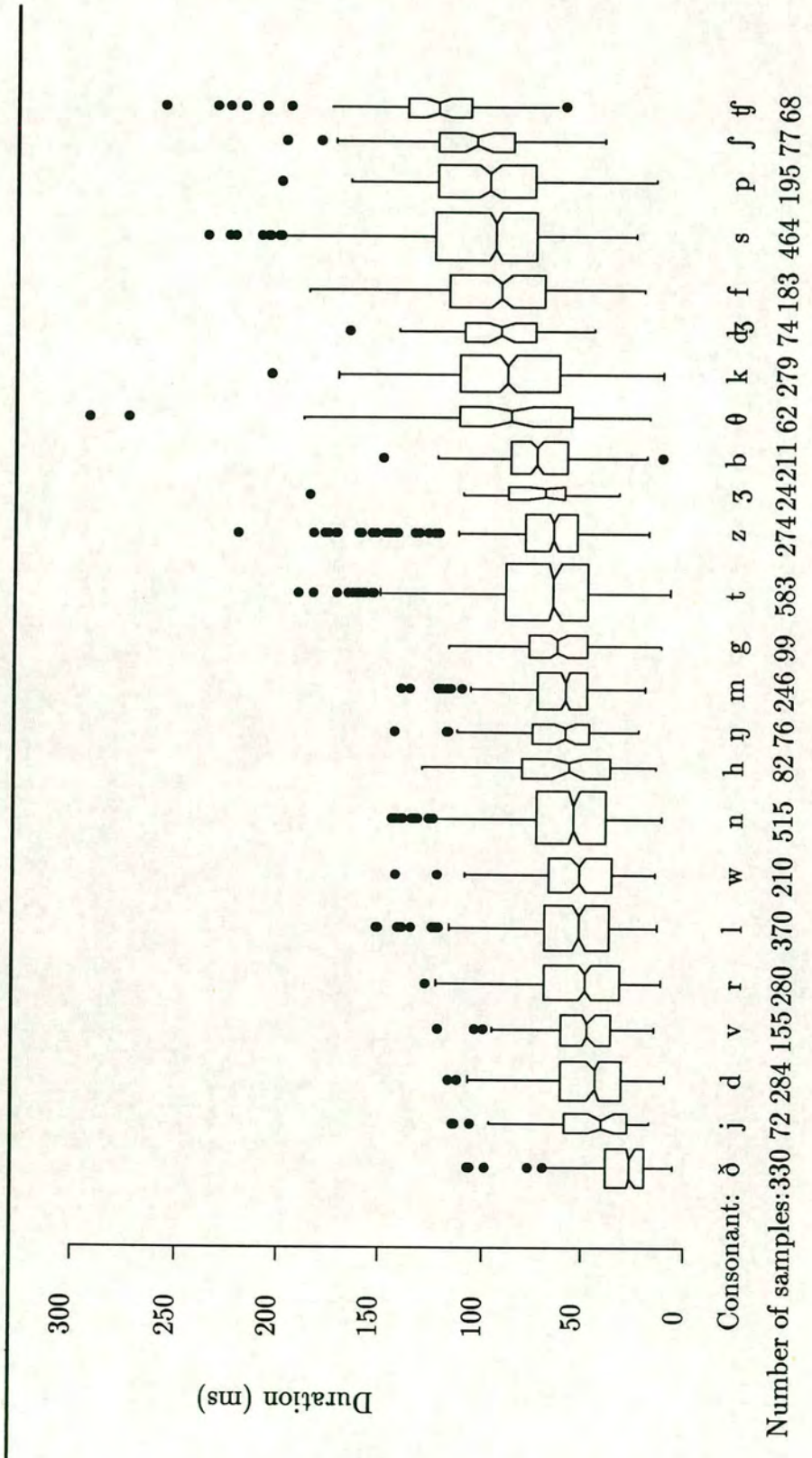


Figure 4.3: Consonant durations

Broad phonetic class	Fine phonetic classes
reduced monophthong	/ə/
short monophthong	/ɪ ʊ ʌ ɛ ɒ a/
long monophthong	/i u ɔ ɑ ɜ/
diphthong	/eɪ ɔɪ aɪ əʊ aʊ ɪə ʊə ɛə/
sonorant	/l r w j n m ŋ/
voiced obstruent	/ð d v g z ʒ b ʤ/
unvoiced obstruent	/θ t f k s ʃ p ʧ h/

Table 4.1: Duration: Broad and fine phonetic classes

basis of phonetic principles (Ladefoged, 1982) and similarities in their observed duration distributions, as indicated in Table 4.1. The effect of using either a fine (P_0) or broad (P_1) phonetic classification on the z-transform normalisation techniques, and subsequently the effect on the prosodic analysis, is investigated.

Each type of unit $U_{1,\ldots,5}$ contains a number of phones and each phone duration can be normalised by a number of different techniques. One of the following measures can be used to represent the duration of a unit in an utterance:

- M_0, unnormalised unit duration, T_{unit}.
- M_1, maximum z_{mean} duration measure of the phones in a unit.
- M_2, maximum $z_{percentile}$ duration measure of the phones in a unit.
- M_3, sum z_{mean} duration measures of all phones in a unit (same as M_1 for unit U_1, since unit U_1 only ever contains one phone).
- M_4, sum $z_{percentile}$ duration measures of all phones in a unit (same as M_2 for unit U_1).

None of the unit duration measures $M_{0,\ldots,4}$ normalise for the number of phones in the unit, as proposed by Fant. Two methods of normalising for the number of phones in the unit are investigated. The first of these methods F_1 is identical to the duration index in Equation 4.1. Therefore, normalisation method F_1 takes into account the variance of unit duration measure as a function of the relative prominence of a unit, as well as taking into

account the number of phones in a unit. This may be redundant for the unit duration measures $M_{2,4}$ based on the $z_{percentile}$-transform. A second normalisation method F_2 is therefore proposed which aims only to compensate for the number of phones in a unit.

- F_0, no normalisation for the number of phones in a unit.
- F_1, duration index for a unit, S_i.

$$S_i = 1 + \frac{M - \overline{M_u}(i)}{\overline{M_s}(i) - \overline{M_u}(i)} \tag{4.4}$$

  where $M \in M_{0,\ldots,4}$, i represents the number of phones in the unit, $\overline{M_u}(i)$ is the mean value of M for all non-prominent units containing i phones, and $\overline{M_s}(i)$ is the mean value of M for all prominent units containing i phones.

- F_2, duration index for a unit, S_i' (equivalent to F_0 for unit-type U_1, since unit U_1 only ever contains one phone).

$$S_i' = M - \overline{M}(i) \tag{4.5}$$

  where $M \in M_{0,\ldots,4}$ and $\overline{M}(i)$ is the mean value of M for all units containing i phones.

The mean unit duration measures $\overline{M_u}(i)$, $\overline{M_s}(i)$ and $\overline{M}(i)$ are determined from the training data. If there are only a few examples of the units containing certain numbers of phones then the sample set can be unrepresentative of the data. Values for the mean unit duration measures are therefore accepted only if there are five or more examples of stressed units and five or more examples of unstressed units containing a particular number of phones, and if $\overline{M_u}(i) < \overline{M_s}(i)$ for that number of phones.

If, in analysing an utterance, a unit is encountered which contains a number of phones for which $\overline{M_u}(i)$, $\overline{M_s}(i)$ and $\overline{M}(i)$ are unknown, then values for these statistics are estimated by applying polynomial interpolation and extrapolation to those which are known. If $\overline{M_u}(i) \geq \overline{M_s}(i)$ for extrapolated values of i, then the normalisation technique breaks down.

In summary, the investigation examines five different types of unit $U_{1,\ldots,5}$. There are five unit duration measures $M_{0,\ldots,4}$ of which only three apply to unit-type U_1. Four of

the unit duration measures $M_{1,\ldots,4}$ are related to the z-transform which can be based on either broad or fine phonetic classes — $P_{0,1}$ — and may or may not involve non-linear smoothing — $S_{0,1}$. Each of the five unit duration measures $M_{0,\ldots,4}$ for unit-types $U_{2,\ldots,5}$ may or may not be processed (in one of two ways) to normalise for the number of phones in the unit — $F_{0,1,2}$. The combinations of different units and normalisation techniques give rise to 213 possible duration features in all, any one of which is an optimal duration correlate of sentential stress.

4.4 Evaluation of duration features

The aim of the investigation is to determine which of these 213 duration features is the most effective measure in distinguishing between non-prominent and prominent syllables. The investigation uses a technique which classifies each syllable as either non-prominent or prominent given a single duration feature. The technique only uses the value of the duration feature for the syllable which is to be classified.

Each syllable is assigned to a category (either as a non-prominent syllable or as a prominent syllable) by a unidimensional Bayesian classifier[2] (see Appendix B) which uses the magnitude of a duration feature as its input parameter. In essence, this technique involves training models which consist of the mean feature vector and the covariance matrix for each category to which a token (in this case, a syllable) can be assigned, and the *a priori* probability of a token existing in a given category. A weighted variance-normalised distance measure (the quadratic discriminant score) is calculated between the feature vector for a given token and the centroid of each trained model. The token is assigned to the category which results in the shortest distance. A by-product of the Bayesian classifier is its *entropy score*, which is the theoretical number of bits of additional information required by the classifier to derive the category of a token (the prominence of a syllable) without error. In comparing the performance of the Bayesian classifier given different duration features, a lower entropy score indicates a better duration feature.

The classification of syllables, using just one duration measure at a time, is applied

[2]The Bayesian classifier is sometimes referred to as Quadratic Discriminant Analysis in the literature.

to a database in an open test. The training data is used to calculate the models for the Bayesian classifier and to determine the mean, percentile and standard deviations of the phone durations. The test data is not used in any of the training procedures.

The entropy scores are determined for the unidimensional Bayesian classifier using each of the 213 duration features as the only input parameter. These scores are shown in Figures 4.4, 4.5, 4.6, 4.7 & 4.8. All entropy scores are shown in each Figure. Each column shows the entropy scores as a function of one variable for each permissible combination of the other four variables. The columns are ranked such that the best combination lies to the left-hand side and the worst combination lies to the right-hand side. For example, each column of Figure 4.4 shows the entropy scores as a function of unit-type $U_{1,\ldots,5}$ for each permissible combination of the variables $M_{0,\ldots,4}$, $P_{0,1}$, $S_{0,1}$ and $F_{0,1,2}$. The combination of phonetic unit and normalisation techniques corresponding to the minimum entropy score (and hence the best combination of those examined) is the syllable lhyme U_3 with sum $z_{percentile}$ duration measures M_4 trained on broad phonetic classes P_1 without smoothing of the phone-level duration contour S_0 and with normalisation of the number of phones in the syllable lhyme using technique F_1.

Figure 4.4 ranks the phonetic units in order of decreasing correlation with sentential stress as the syllable lhyme U_3, the syllable nucleus U_1, the entire syllable U_4, the syllable rhyme U_2 and finally the nucleus-to-nucleus unit U_5. The correlation with syllable prominence therefore decreases with an increase in the number of phones to the right of the syllable nucleus relative to the number of phones to the left of the nucleus. Factors other than prominence cause the duration of phones to the right of the nucleus to vary more than the duration of phones to the left of the nucleus. This result supports rules for the synthesis of segment durations proposed by Klatt (1979). Klatt's rules shorten the duration of consonants in a non-word-initial position (phones in the syllable coda) and lengthen the duration of the syllable nucleus and any following consonants (phones in the rhyme) when they are in a pre-pausal position. It is claimed by Campbell (1990) that duration lengthening due to prominence affects the entire syllable U_4 whereas lengthening due to prosodic boundaries affects the syllable rhyme U_2. Wightman & Ostendorf (1991; 1992) use the average z_{mean} duration measure of phones in the rhyme of the final syllable in a word as a feature for the automatic assignment of break indices. The break indices

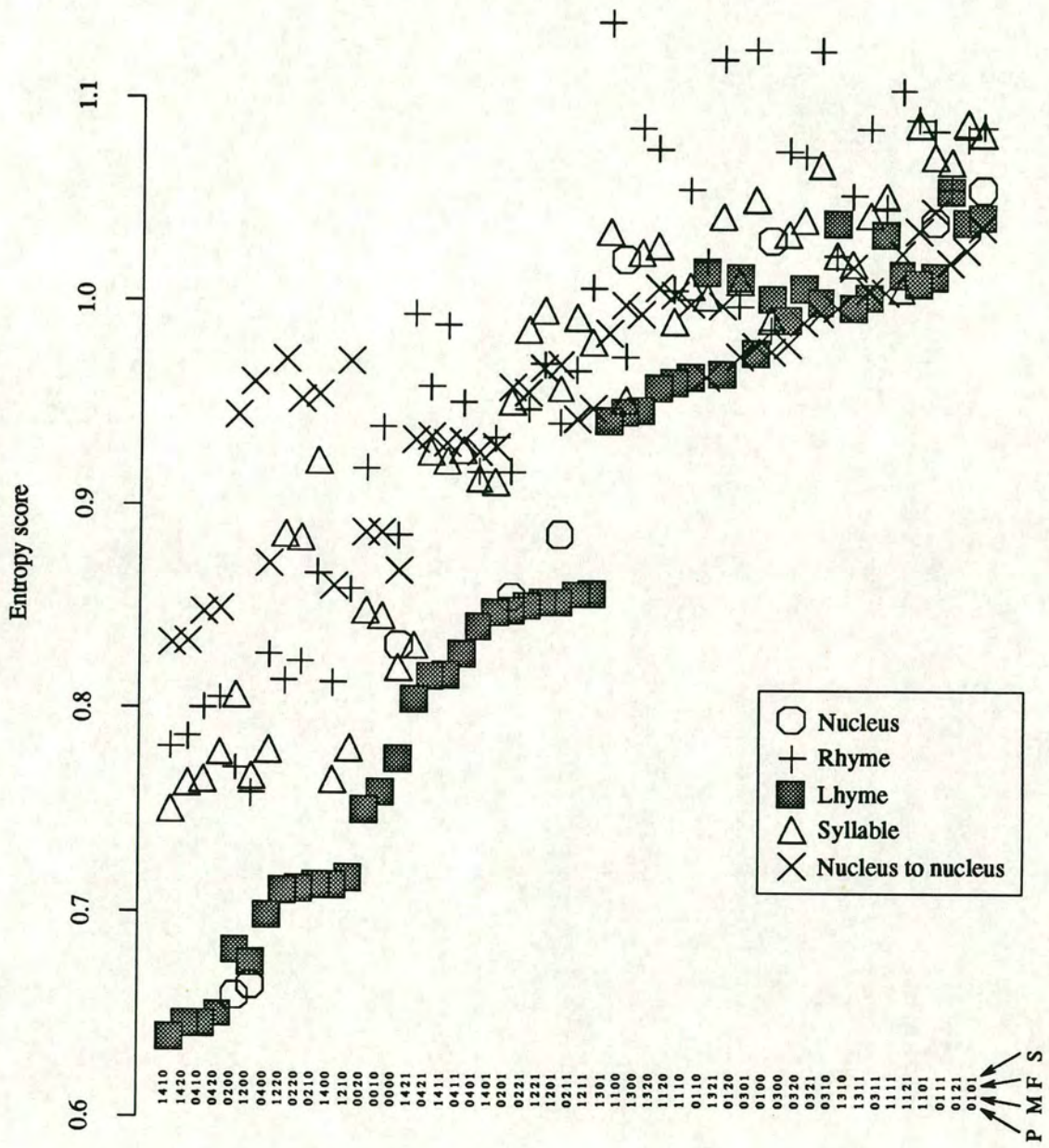


Figure 4.4: Duration feature evaluation: Entropy score ($U_{1,\ldots,5}$)

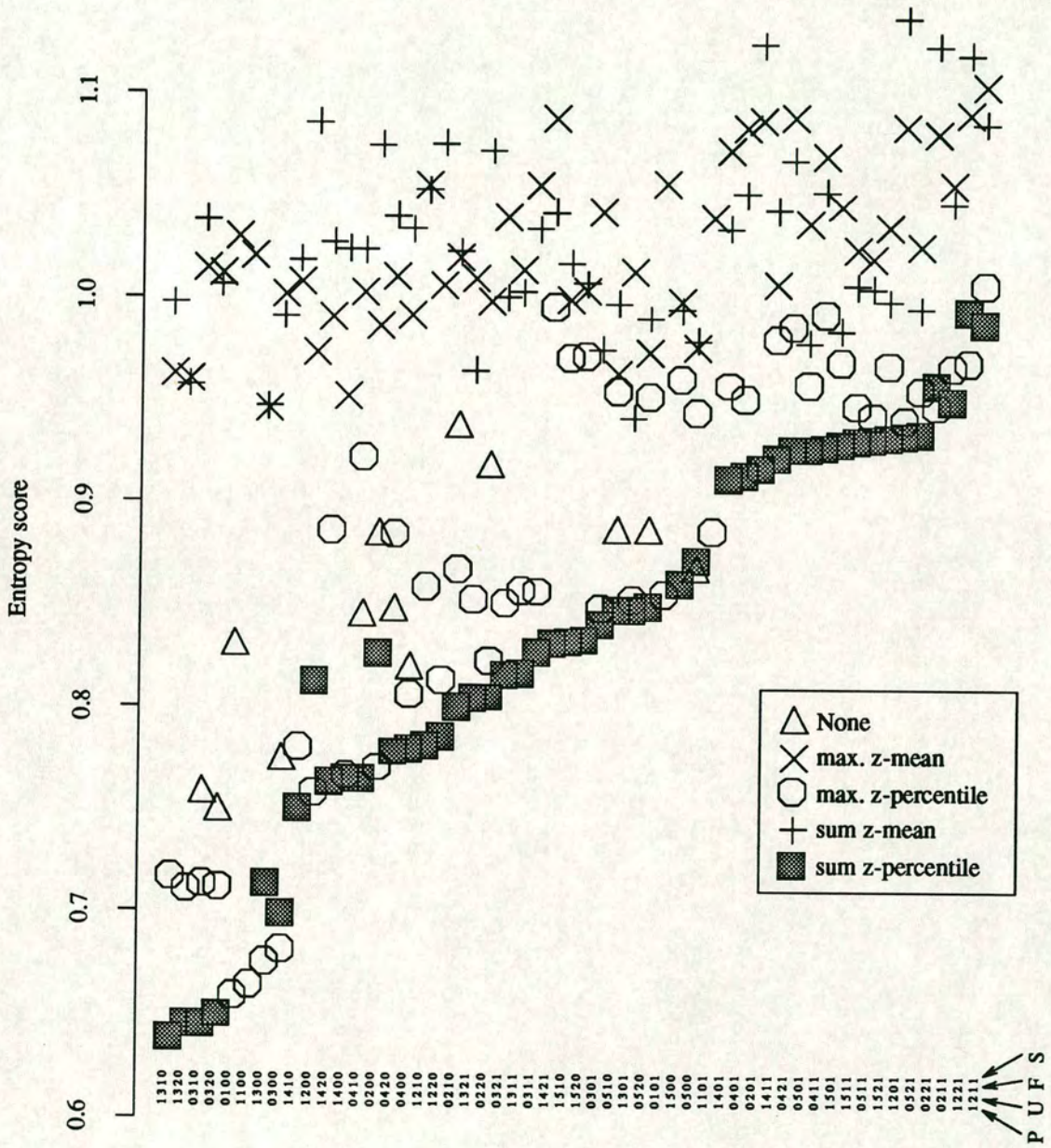


Figure 4.5: Duration feature evaluation: Entropy score ($M_{0,\ldots,4}$)

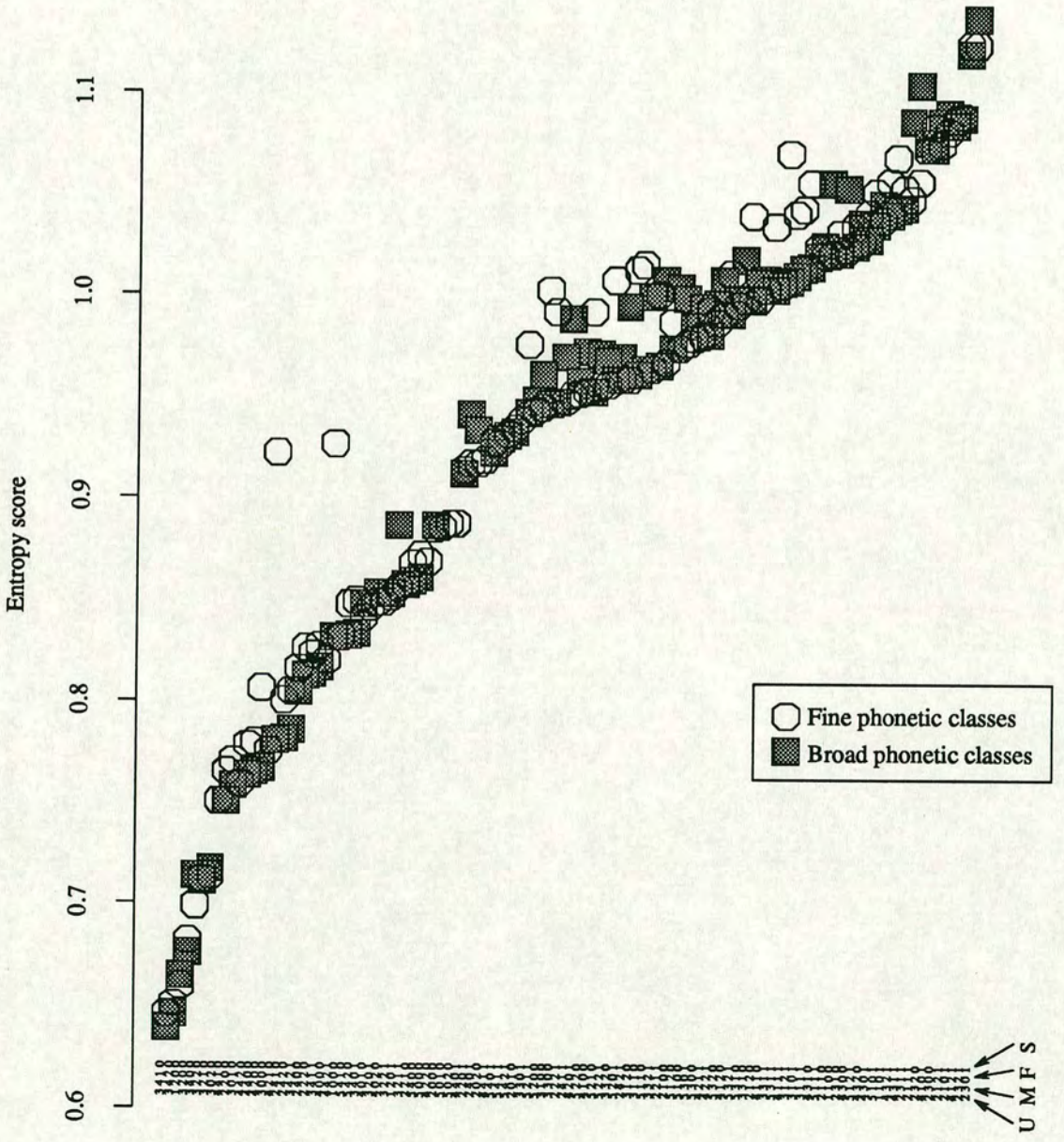


Figure 4.6: Duration feature evaluation: Entropy score ($P_{0,1}$)

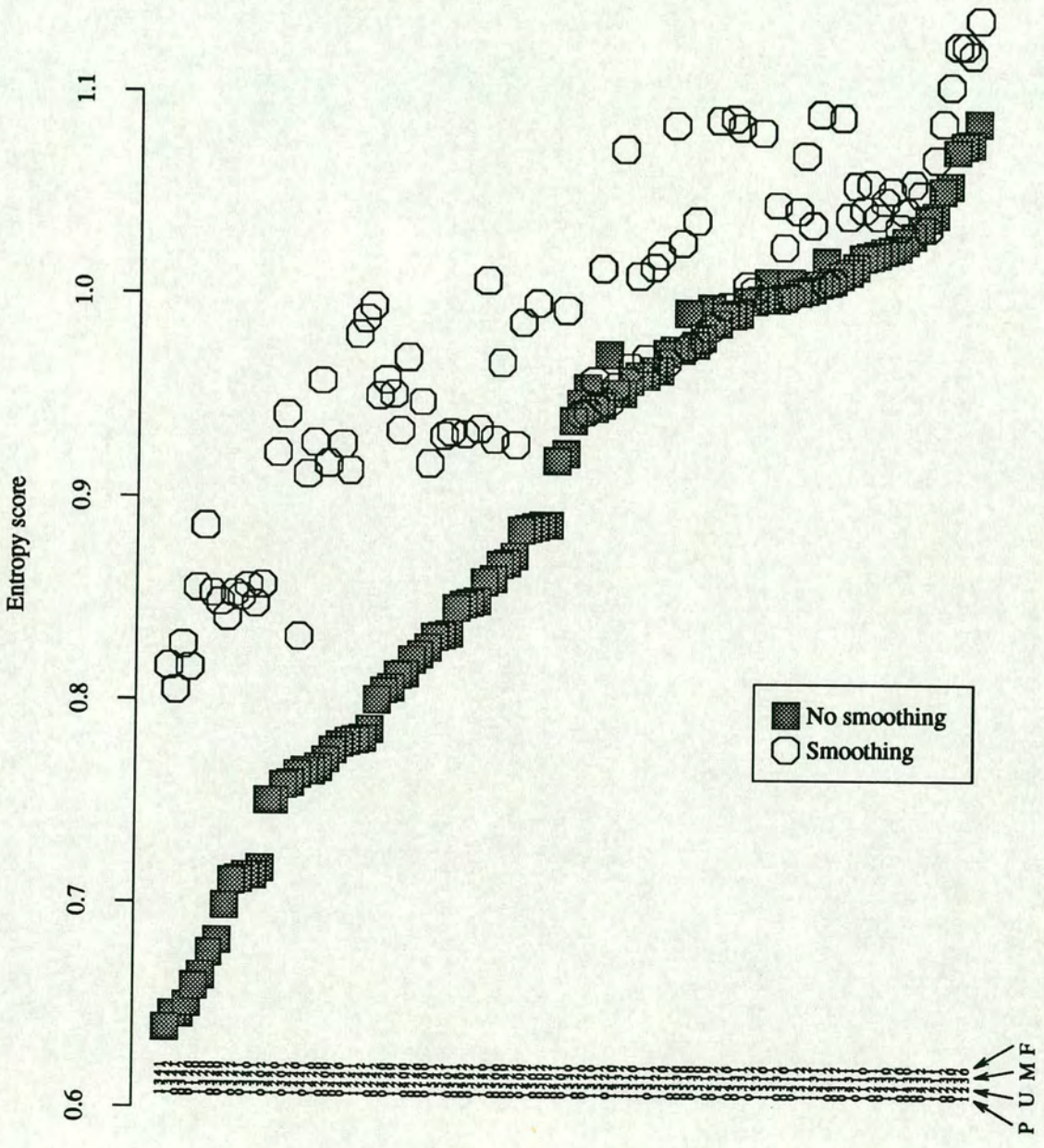


Figure 4.7: Duration feature evaluation: Entropy score ($S_{0,1}$)

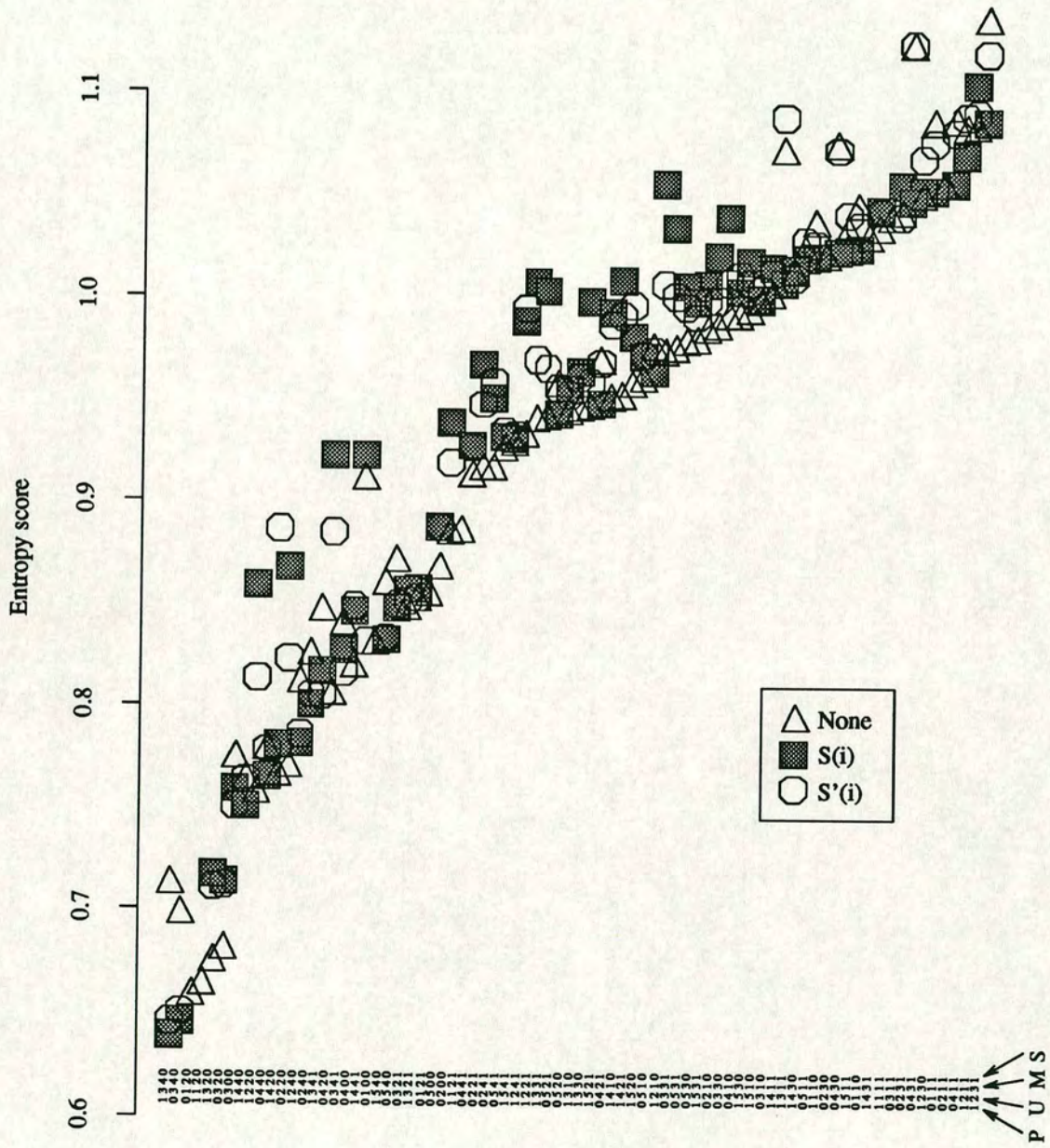


Figure 4.8: Duration feature evaluation: Entropy score ($F_{0,1,2}$)

are used to represent the strength of boundaries between neighbouring words and are thus related to the prosodic structure of an utterance.

Figure 4.5 shows a clear division between z_{mean} based normalisations (represented by plusses and crosses) and $z_{percentile}$ based normalisations (represented by octagons and shaded squares). The $z_{percentile}$ duration measures yield the preferable lower entropy scores. Lower entropy scores are obtained by using the sum of z-transformed phone durations in a unit than are obtained by using the maximum z-transformed phone duration. The z_{mean} duration measures give consistently higher entropy scores than using no normalisation for phone-type.

There is little difference in the entropy scores shown in Figure 4.6 for z-transformation statistics trained on fine and broad phonetic classes. Normalisation for broad phonetic classes results in a slightly lower entropy score than normalisation for fine phonetic classes. There is therefore only a small possibility that the training data is unrepresentative of the universal set of phones classified on a fine phonetic basis (with respect to duration analysis).

Smoothing of the duration contour causes an increase in entropy, as illustrated in Figure 4.7. The smoothing process therefore eliminates information required in the prosodic analysis rather than compensating for phone boundary placement errors in a way which could aid the analysis. The application of such smoothing is therefore not beneficial to the prosodic analysis.

Figure 4.8 reveals that normalisation for the number of phones in a unit is dependent upon the duration measure and phonetic unit-type used. No normalisation for the number of phones F_0 gives a lower entropy score for many of the combinations of duration measure and phonetic unit-type. However, in the case of the sum $z_{percentile}$ duration measure M_4 for the lhyme U_3 (shown in the left-most column) the entropy score is significantly lower when normalisation for the number of phones in the lhyme is incorporated. Normalisation method F_1 takes into account the variance of the lhyme duration as a function of the relative prominence of the syllable, as well as taking into account the number of phones in the lhyme. A degree of redundancy is expected for the $z_{percentile}$-based duration measures $M_{2,4}$ since the $z_{percentile}$-transform also takes into account the variance of the lhyme duration as a function of the relative prominence of

the syllable. Normalisation method F_2 only compensates for the number of phones in the lhyme. Figure 4.8 shows that the difference in entropy score between methods F_1 and F_2 is small (but statistically significant), with method F_1 resulting in the lower entropy score. A difference in entropy exists because the assumptions made about the nature of the variation in the lhyme duration due to the prominence of a syllable differ in the F_1 normalisation method and the $z_{percentile}$-transform. Method F_1 assumes that the mean duration measure for prominent syllables is greater than the mean duration measure for non-prominent syllables. The $z_{percentile}$-transform assumes that all of the lower duration measures correspond to non-prominent syllables and that all of the higher duration measures correspond to prominent syllables. Making both assumptions yields a lower entropy score.

4.5 Optimisation of a relative duration feature

The application of a Bayesian classifier assumes that the prominence of a syllable can be determined from the magnitude of a duration feature in isolation from its neighbours. The variation of a duration feature relative to neighbouring syllables is not included as an input parameter to the classifier. This exclusion cannot be beneficial to the Bayesian classifier, given that a syllable's prominence is relative to its neighbours by definition. A possible solution to this problem is to include the duration measures for neighbouring syllables as input parameters to the Bayesian classifier. This might improve the Bayesian classifier's performance (in terms of reducing the entropy score) when given the task of classifying a syllable as either non-prominent or prominent. However, this approach is not investigated here because an emphasis is placed on employing non-statistical methods such that the underlying principles used to locate prominent syllables can be easily identified.

A simple "peak-picking" technique is proposed here which classifies the syllable as either non-prominent or prominent depending on the magnitude of the duration feature relative to its nearest neighbours. A syllable is classified as prominent if its duration feature exceeds that of both its neighbours (end-points being inherently lower) and if such a local maximum in the duration contour has a magnitude greater than some specified threshold $K_{duration}$. Syllables satisfying this criterion are labelled "*sd*". A syllable is

also classified as prominent if it corresponds with the maximum value of the duration feature in the contour (independently of the threshold). Syllables corresponding to such maxima are labelled "*Sd*". All other syllables are labelled "*ud*".

The duration feature used corresponds to the syllable lhyme U_3 with sum $z_{percentile}$ durations M_4 trained on broad phonetic classes P_1 without smoothing of the phone-level duration contour S_0 and with normalisation of the number of phones in the syllable lhyme using technique F_1 — ie. the duration feature with the minimum entropy score in the above investigation. Theoretically, this duration feature has a mean value of 2.0 for prominent syllables and a mean value of 1.0 for non-prominent syllables. $K_{duration}$ should therefore be optimal at 1.5. The performance of the peak-picking method is evaluated by examining the percentage of syllables which are assigned to the correct category[3] (prominent or non-prominent). The optimal value of $K_{duration}$ (ie. the value corresponding to the greatest percentage of syllables assigned to the correct prominence category) is sought in a closed test. The data used to train the $z_{percentile}$ statistics and the mean duration measures used in the F_1 normalisation method is also used in optimising $K_{duration}$. A plot of the percentage of correct categorisation against $K_{duration}$ is shown in Figure 4.9. The optimal value of $K_{duration}$ is 1.5, as predicted.

An entropy score is not produced by the peak-picking method. Its performance is therefore compared with the performance of the Bayesian classifier by examining the percentage of syllables which are assigned to the correct prominence category in an open test. The levels of agreement between the prominence category of a syllable as transcribed by hand and as assigned automatically are shown in Tables 4.2 & 4.3 for the peak-picking algorithm and for the Bayesian classifier respectively. The simple peak-picking algorithm has a number of benefits over the Bayesian classifier. The peak-picking algorithm yields an improvement in syllable prominence classification over the Bayesian classifier, requires less computation (and hence it is faster) than the Bayesian classifier, and does not requiring a training stage.

[3]It is assumed that the "correct" prominence category is that which is transcribed by hand, despite the possibility of human error.

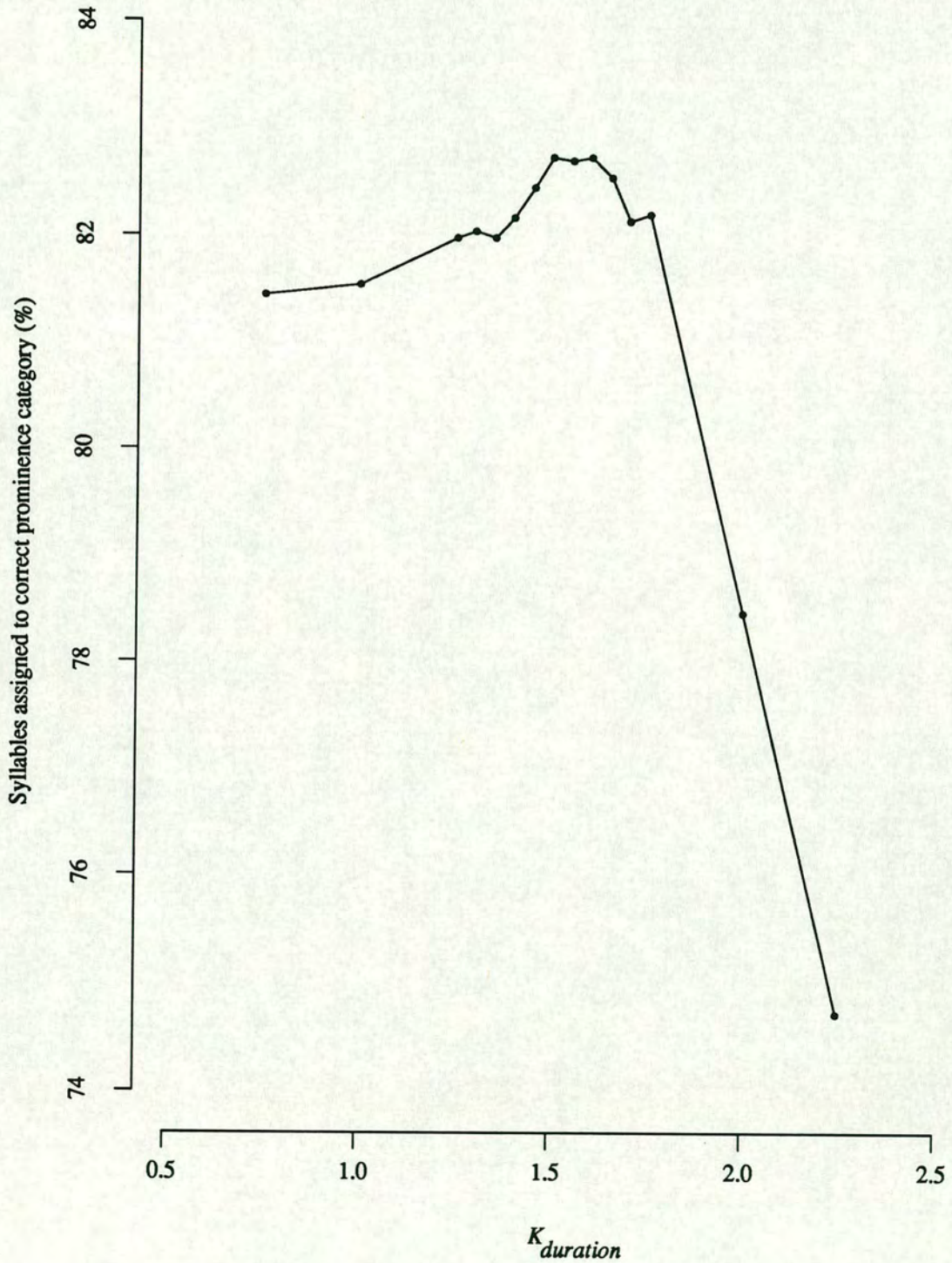


Figure 4.9: Optimisation of $K_{duration}$ threshold

		Peak-picking algorithm label			
		Sd	*sd*	*ud*	total
Hand Label	*P*	367 (6.9%)	1027 (19.3%)	704 (13.2%)	2098 (39.4%)
	NP	93 (1.7%)	386 (7.3%)	2742 (51.6%)	3221 (60.6%)
	total	460 (8.6%)	1413 (26.6%)	3446 (64.8%)	5319 (100.0%)

Correct classification rate = 4136/5319 (77.8%)

Table 4.2: Duration: Confusion matrix for peak-picking algorithm
(*Sd* — syllable with maximum duration of contour; *sd* — syllable corresponding to a local peak in the duration contour greater than $K_{duration}$; *ud* — unstressed syllable on the basis of duration; *P* — prominent syllable; *NP* — non-prominent syllable)

		Bayesian classifier label		
		P	*NP*	total
Hand Label	*P*	1565 (29.4%)	533 (10.0%)	2098 (39.4%)
	NP	723 (13.6%)	2498 (47.0%)	3221 (60.6%)
	total	2288 (43.0%)	3031 (57.0%)	5319 (100.0%)

Correct classification rate = 4063/5319 (76.4%)

Table 4.3: Duration: Confusion matrix for Bayesian classification
(*P* — prominent syllable; *NP* — non-prominent syllable)

4.6 Summary

The investigation described above aimed to determine a duration feature whose distributions for prominent and non-prominent syllables have the greatest separability. The entropy score obtained by a Bayesian classifier was used as a robust measure of this separability for 213 different duration features. The optimal duration feature $D_{feature}$ of those investigated is given by:

$$D_{feature} = 1 + \frac{M - \overline{M_u}(i)}{\overline{M_s}(i) - \overline{M_u}(i)} \tag{4.6}$$

where,

$$M = \sum_{phone \in lhyme} \frac{T_{phone} - \varphi_T^p(broad\_phone\_class)}{\sigma_T(broad\_phone\_class)} \tag{4.7}$$

and where i represents the number of phones in the lhyme, $\overline{M_u}(i)$ is the mean value of M for all non-prominent lhymes containing i phones, $\overline{M_s}(i)$ is the mean value of M for all prominent lhymes containing i phones in the training data, T_{phone} is the duration of a phone, $\varphi_T^p(broad\_phone\_class)$ and $\sigma_T(broad\_phone\_class)$ are the p'th percentile and the standard deviation of phone durations grouped on a broad phonetic basis respectively, and p is the percentage of non-prominent phones in the training data.

A simple peak-picking algorithm applied to this duration feature is shown to give an improvement of syllable prominence classification over the Bayesian classifier, while requiring less computation and not requiring an additional training stage. In an open test, 77.8% of syllables are given the same prominence status by the peak-picking algorithm as when transcribed by hand. This level of agreement is dependent upon the discrete categorisation of syllable prominence perceived by a human transcriber, and as such, is regarded in this thesis as a comparative measure only.

The duration features investigated include the normalisation of variations in duration due to phone-type, syllable structure, syllable length in terms of number of phones, and the prominence level of syllables within the training data. However, these are not the only factors which influence the duration of a segment. The normalisation of segmental context effects and pre-pausal lengthening are not included in the investigation (although it is argued that the lhyme is not affected by pre-pausal lengthening to as great an extent as other syllable components). Further research is required to investigate the possible advantages that can be gained by incorporating models of contextual influences on segment duration (Bartkova & Sorin, 1987; Crystal & House, 1988; van Santen, 1993) into the normalisation of non-prosodic effects. The duration of speech segments is also affected by variations in the rate of articulation from utterance to utterance. The inclusion of compensating for articulation rate differences (Wightman *et al.*, 1992) is also a subject of further investigation.

Chapter 5

Energy measures

The formation of a frame-level low-band energy contour is described in Section 5.1. The contour is used as an acoustic representation of the perceived variations of intensity in an utterance. The frame-level low-band energy contour is reduced to a phone-level low-band energy contour so that each phone in an utterance is associated with a single measure of intensity. The low-band energy of phones in a variety of phonetic units (Section 4.2) and normalisation techniques are investigated in this Chapter with respect to forming an energy feature correlated with sentential stress. The methodology used in the investigation of duration features is also used in this Chapter.

5.1 Low-band energy contour

The energy contour for a speech waveform is calculated from frames of data (at 5ms intervals) such that energy values are synchronised with the data used in *FØ* extraction (see Chapter 7) and with the formant frequency calculations used in the vowel quality analysis (see Chapter 6). Each frame is passed through a Blackman-Harris window (Harris, 1978) in order to filter out discontinuities (and hence high frequency artefacts) at the boundaries of the analysis frame. The frequency bins of an amplitude spectrum (calculated by a 512-point FFT for 20ms frames of data sampled at 20kHz) corresponding to the range 50Hz–2kHz are accumulated. The resultant energy values are expressed in decibels with respect to the maximum frame energy in the utterance to form an utterance-normalised low-band energy contour. Expressing the energy values in these

terms normalises for utterance-to-utterance variations in energy which are dependent upon recording conditions such as the extent of signal pre-amplification and the distance of the microphone from the voice source. The contour is processed by a three-frame median filter and five-frame Hann window smoother (Rabiner *et al.*, 1975) in order to remove small perturbations which arise during frames of speech with low fundamental frequency (typically less than two fundamental periods per analysis frame).

A phone-level low-band energy contour is generated from the (frame-level) utterance-normalised energy contour. Each phone in the phonetic transcription of an utterance is associated with the energy value of one of the frames within the phone. If one or more peaks (local maxima flanked by lower values) in the energy contour occur within the phone then the energy value at the highest peak is associated with the phone. If such a peak does not exist then either a single valley in the energy contour occurs within the phone or the energy contour continuously rises or falls during the phone. If a single valley exists then the energy value at the valley is associated with the phone. Otherwise, the phone is associated with the energy value at its mid-point.

An example of a phone-level low-band energy contour and its corresponding (frame-level) utterance-normalised energy contour and speech waveform are illustrated in Figure 5.1. The dots on the frame-level energy contour indicate the frame energy peak/valley/mid-point values associated with each of the phones. The phone-level contour follows the same general trends as the frame-level contour. By associating a phone with the frame energy value at the point of a local maximum flanked by lower values during the phone rather than with the maximum frame energy value in the phone, the formation of the phone-level low-band energy contour becomes more robust to variations in the placement of phone boundaries. This is illustrated in the cases of the second [f] and the [v] in Figure 5.1. The peak/valley/mid-point value associated with a phone is also robust to phonetic-context. For example, the value of the valley in the [t] is less changeable to variations in the peak energies of the neighbouring vowels than the maximum frame energy of the phone. Although the peak/valley/mid-point value is robust to phone boundary placement and phonetic context, it is not immune to potential errors. If the boundary between the [i] and the [l] is placed beyond the local valley, or if the left context of the [l] is a low low-band energy phone such as an /f/, then the [l] would be

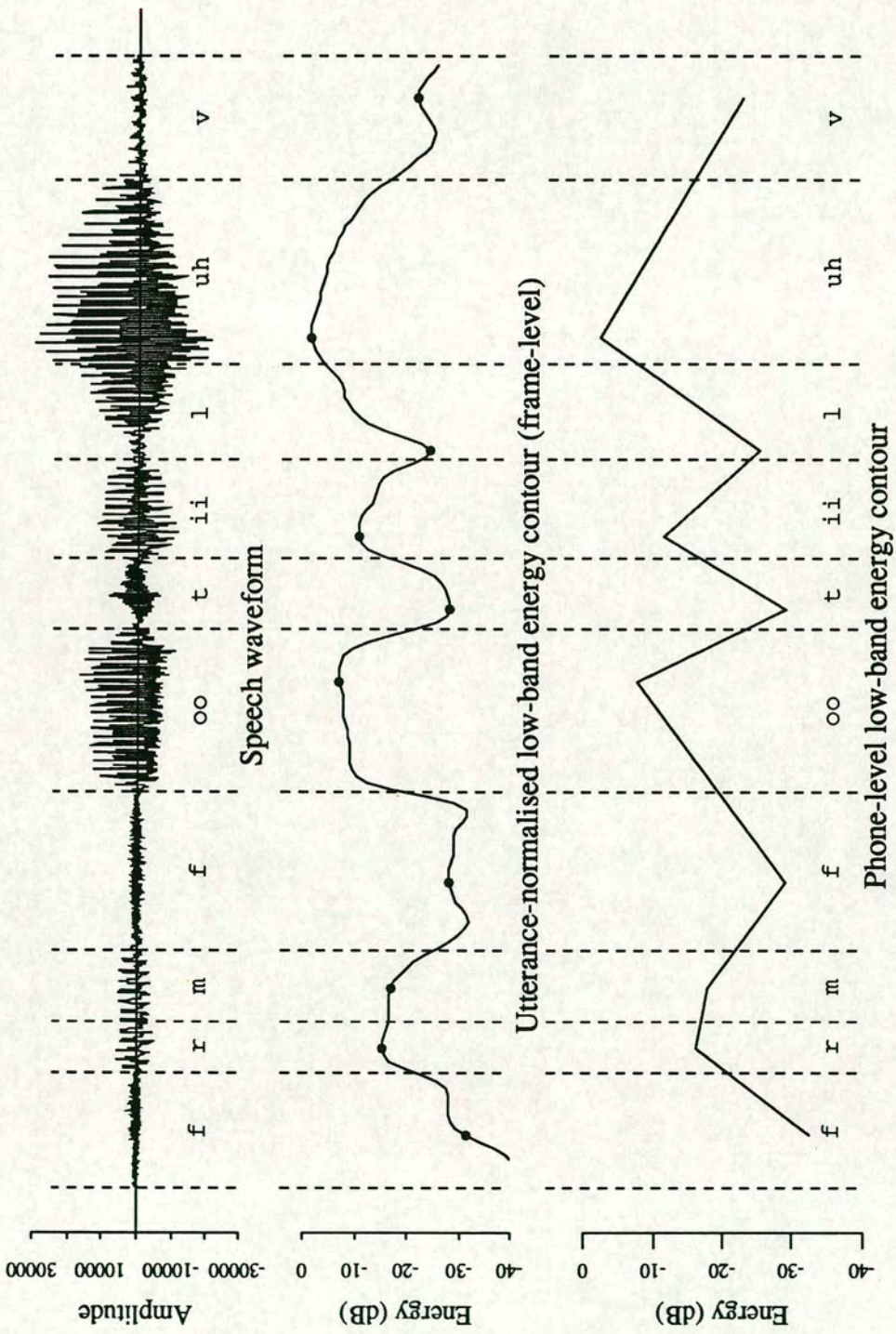


Figure 5.1: Formation of a phone-level low-band energy contour*

associated with the higher mid-point frame energy value rather than the frame energy value at the local valley.

5.2 Energy features

The low-band energy of each phone in an utterance is calculated using the procedure described in Section 5.1. The phonetic units whose low-band energy best correlate with sentential stress and normalisation techniques which can be applied to them, are studied here. As in the investigation of duration (Chapter 4), the aim is to determine an energy feature whose distributions for each prominence level have the greatest separability.

The units whose energies are investigated are the same as those used in the duration analysis, $U_{1,\ldots,5}$ (Section 4.2). The energy of a unit, like the duration of a unit, is influenced by many parameters other than just its relative prominence. Normalisation techniques are adopted from the duration analysis (Section 4.3) to compensate for any variations in the energy of a unit that result from such parameters.

One of the following measures can be used to represent the low-band energy of a unit in an utterance:

- M_0, unnormalised unit-level low-band energy, E_{unit}. The unit-level low-band energy is generated from the (frame-level) utterance-normalised energy contour by the same peak/valley/mid-point procedure used to determine the phone-level low-band energy.

- M_1, maximum z_{mean} low-band energy measure of the phones in a unit.

$$z_{mean} = \frac{E_{phone} - \mu_E(phone\_type)}{\sigma_E(phone\_type)} \tag{5.1}$$

  where $\mu_E(phone\_type)$ and $\sigma_E(phone\_type)$ represent the mean and standard deviation phone-level low-band energy respectively for each phone-type.

- M_2, maximum $z_{percentile}$ low-band energy measure of the phones in a unit.

$$z_{percentile} = \frac{E_{phone} - \varphi^p_E(phone)\_type}{\sigma_E(phone\_type)} \tag{5.2}$$

Broad phonetic class	Fine phonetic class
closed vowel	/ɪ i ʊ u/
half-closed vowel	/ʊə ɪə eɪ əʊ ɔɪ/
central vowel	/ə/
half-open vowel	/ɔ aɪ ɛ ɛə aʊ ʌ ɜ/
open vowel	/ɒ a ɑ/
liquid	/l r/
glide	/w j/
nasal	/n m ŋ/
voiced obstruent	/b ʒ v g ʤ ð d z/
unvoiced obstruent	/s θ t ʧ f ʃ p k h/

Table 5.1: Low-band energy: Broad and fine phonetic classes

where $\varphi_E^p(phone\_type)$ is the p'th-percentile phone-level low-band energy for each phone-type, and p is the percentage of a phone-type which are non-prominent.

- M_3, sum z_{mean} low-band energy measures of all phones in a unit (same as M_1 for unit U_1, since unit U_1 only ever contains one phone).

- M_4, sum $z_{percentile}$ low-band energy measures of all phones in a unit (same as M_2 for unit U_1).

As in the duration analysis, phone-types can be classified on either a fine or a broad phonetic basis for the z-transform normalisation of phone-level low-band energies. The low-band energy distributions for phones classified on a fine phonetic basis are shown in Figures 5.2 (vowels) & 5.3 (consonants). Phones may be grouped into broad phonetic classes on the basis of phonetic principles (Ladefoged, 1982) and similarities in their observed low-band energy distributions, as indicated in Table 5.1.

Each of the unit low-band energy measures can be normalised for the number of phones in the unit by the techniques $F_{0,1,2}$ described in Section 4.3.

In summary, the investigation examines five different types of unit $U_{1,\ldots,5}$. There are five unit low-band energy measures $M_{0,\ldots,4}$ of which only three apply to unit-type U_1. Four of the unit low-band energy measures $M_{1,\ldots,4}$ are related to the z-transform which can be based on either broad or fine phonetic classes — $P_{0,1}$. No smoothing is

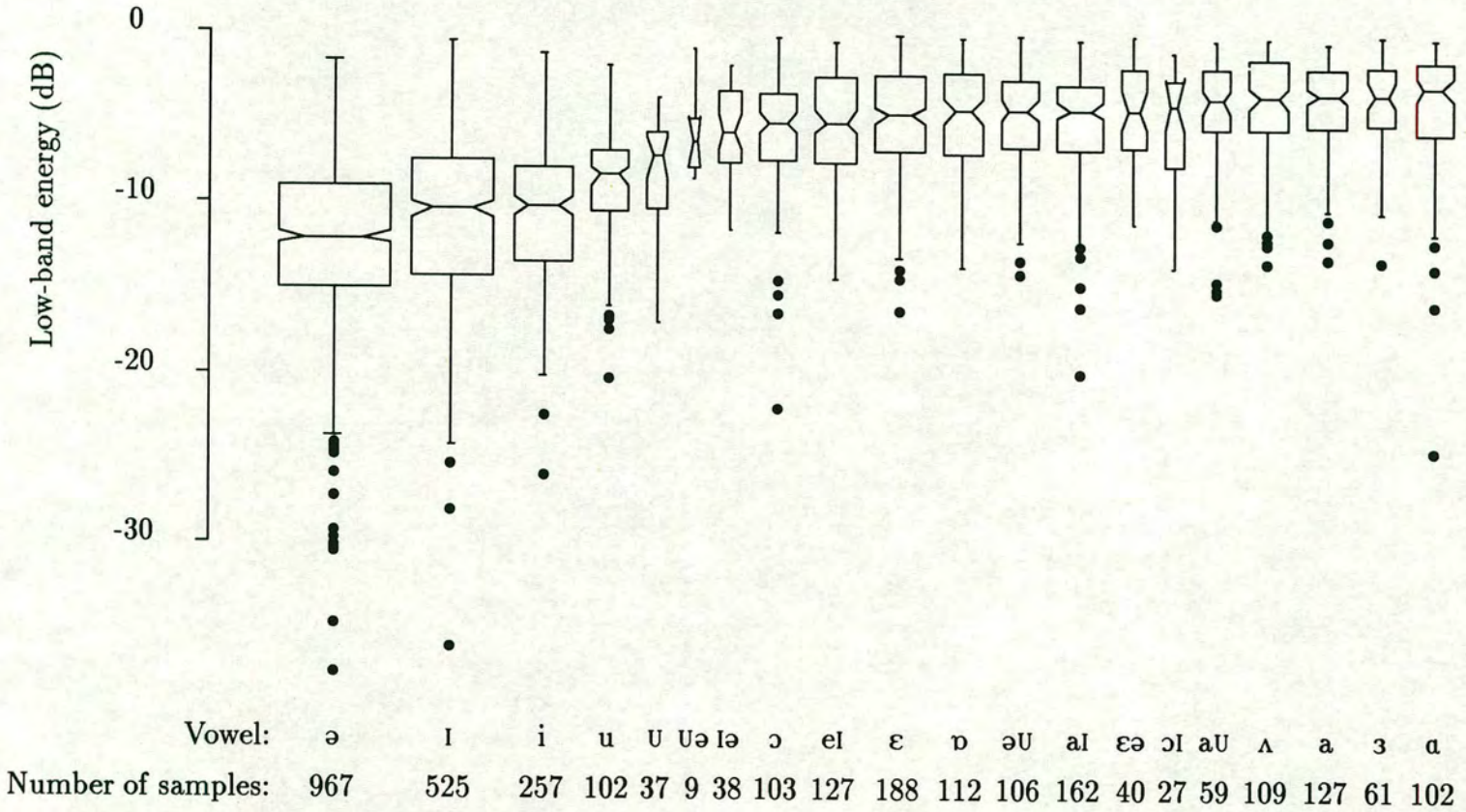


Figure 5.2: Vowel low-band energies

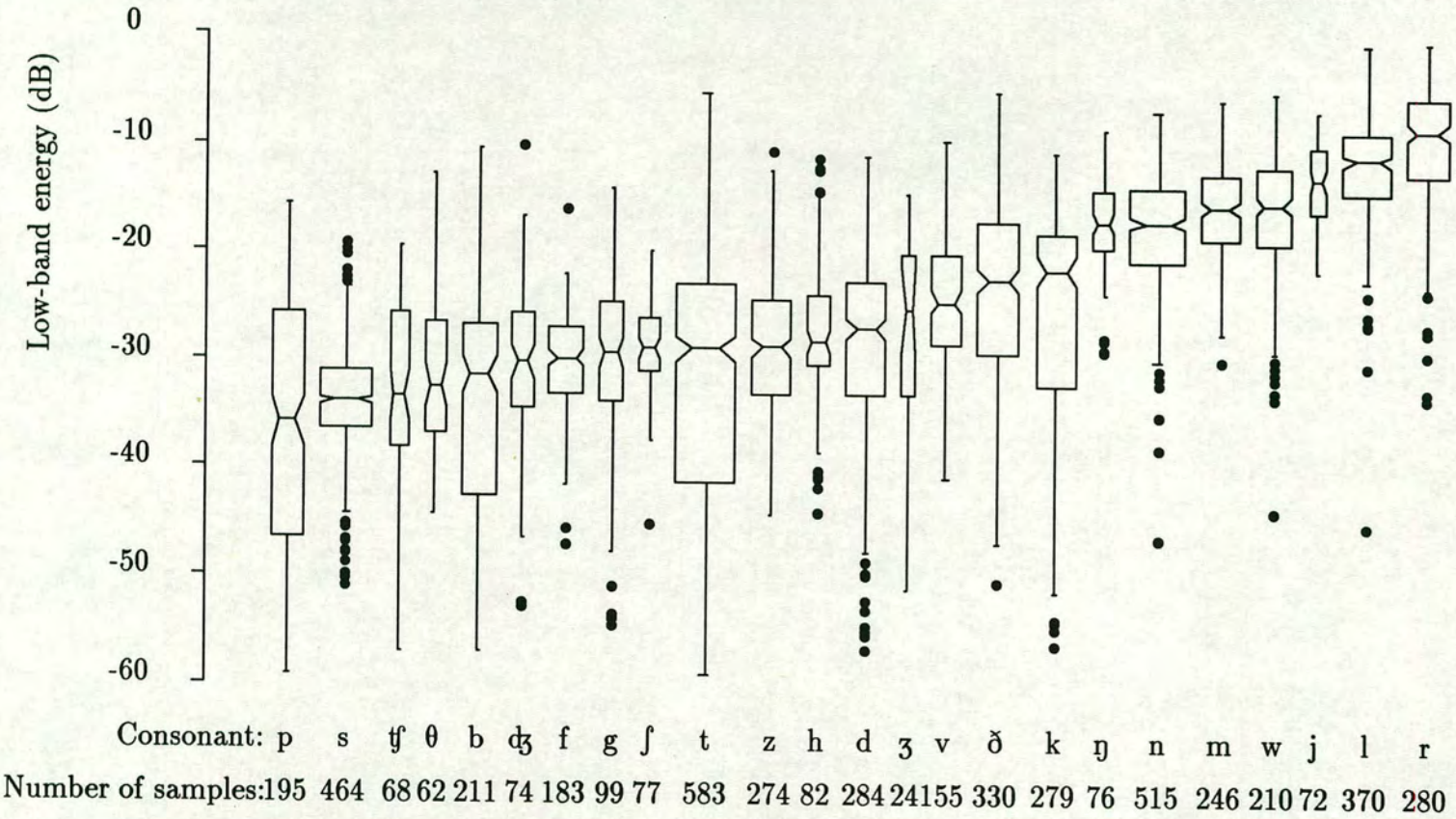


Figure 5.3: Consonant low-band energies

applied to the phone-level low-band energy contour. (Smoothing is applied to the frame-level low-band energy contour.) Each of the five unit low-band energy measures $M_{0,\ldots,4}$ for unit-types $U_{2,\ldots,5}$ may or may not be processed (in one of two ways) to normalise for the number of phones in the unit — $F_{0,1,2}$. The combinations of different units and normalisation techniques give rise to 113 possible energy features in all, any one of which is an optimal energy correlate of sentential stress.

5.3 Evaluation of energy features

The aim of the investigation is to determine which of these 113 energy features is the most effective measure in distinguishing between non-prominent and prominent syllables.

The entropy scores are determined for the unidimensional Bayesian classifier using each of the 113 energy features as the sole input parameter. These scores are shown in Figures 5.4, 5.5, 5.6 & 5.7. All entropy scores are shown in each Figure. Each column shows the entropy scores as a function of one variable for each permissible combination of the other three variables. The columns are ranked such that the best combination lies to the left-hand side and the worst combination lies to the right-hand side. For example, each column of Figure 5.4 shows the entropy scores as a function of unit-type $U_{1,\ldots,5}$ for each permissible combination of the variables $M_{0,\ldots,4}$, $P_{0,1}$ and $F_{0,1,2}$. The combination of phonetic unit and normalisation techniques corresponding to the minimum entropy score (and hence the best combination of those examined with respect to energy) is the syllable nucleus U_1 with $z_{percentile}$ low-band energy measures M_2 trained on fine phonetic classes P_0. Normalisation for the number of phones in a unit $F_{1,2}$ are not applicable to the syllable nucleus.

Figure 5.4 ranks the phonetic units in order of decreasing correlation with sentential stress as the syllable nucleus U_1, the syllable rhyme U_2, the entire syllable U_4, the syllable lhyme U_3 and finally the nucleus-to-nucleus unit U_5. By definition, the syllable nucleus is the peak of sonority in a syllable. Sentential stress is therefore correlated with an energy feature based on the peak of sonority in a syllable.

As is observed in the investigation of duration features, Figure 5.5 shows a clear division between z_{mean} based normalisations and $z_{percentile}$ based normalisations, in this case, with respect to energy. Lower entropy scores are yielded by $z_{percentile}$ low-band

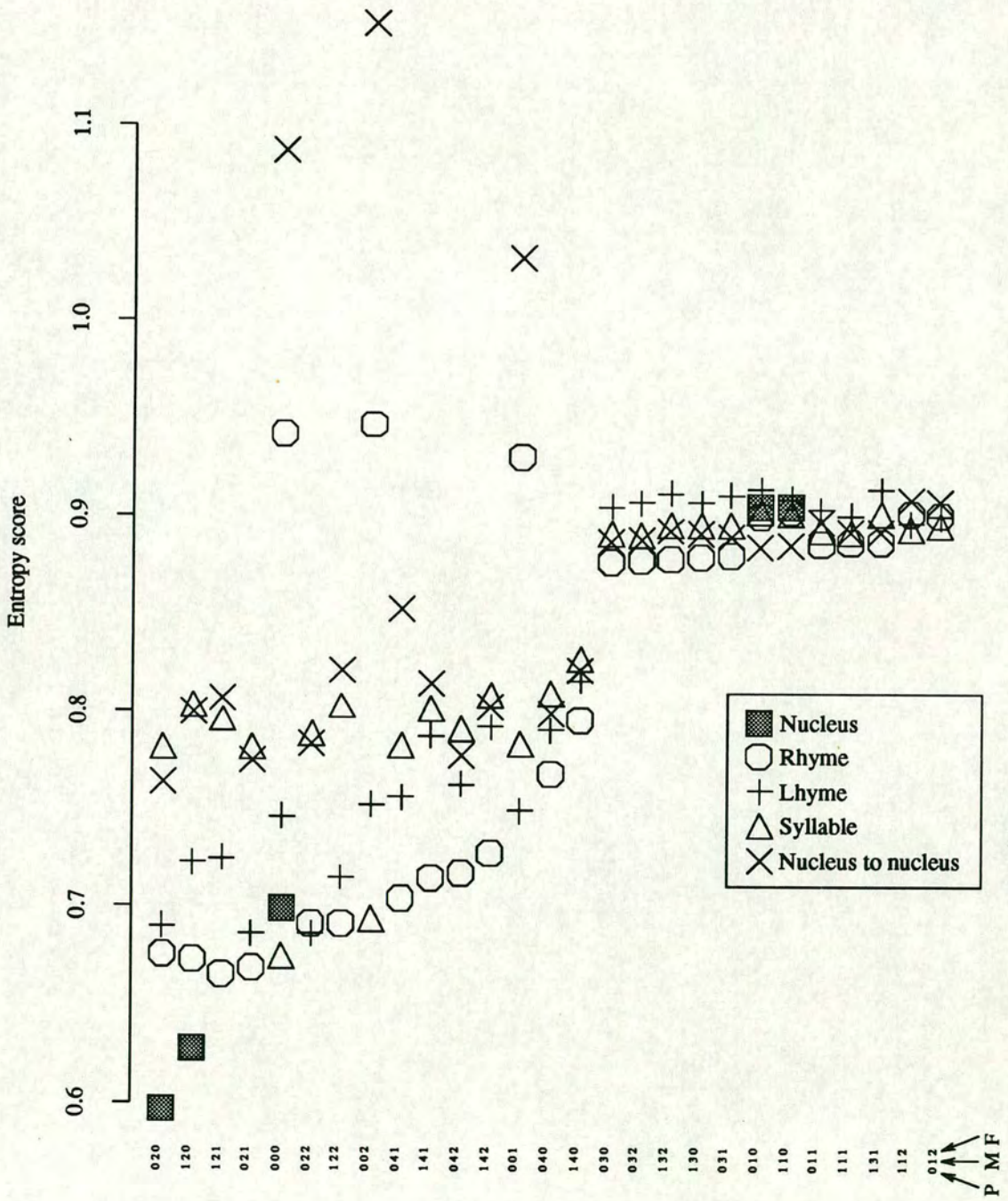


Figure 5.4: Energy feature evaluation: Entropy score ($U_{1,\ldots,5}$)

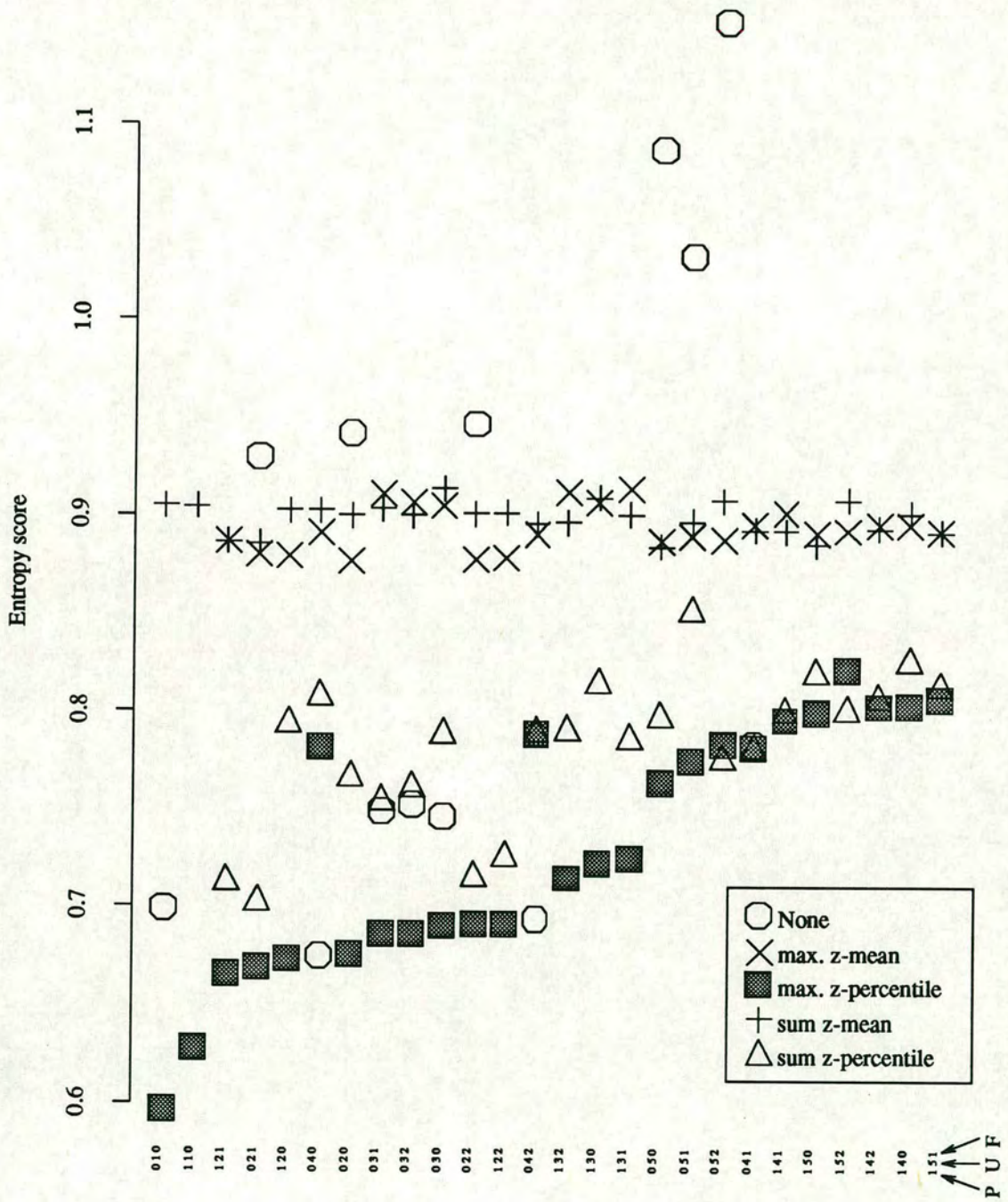


Figure 5.5: Energy feature evaluation: Entropy score ($M_{0,\ldots,4}$)

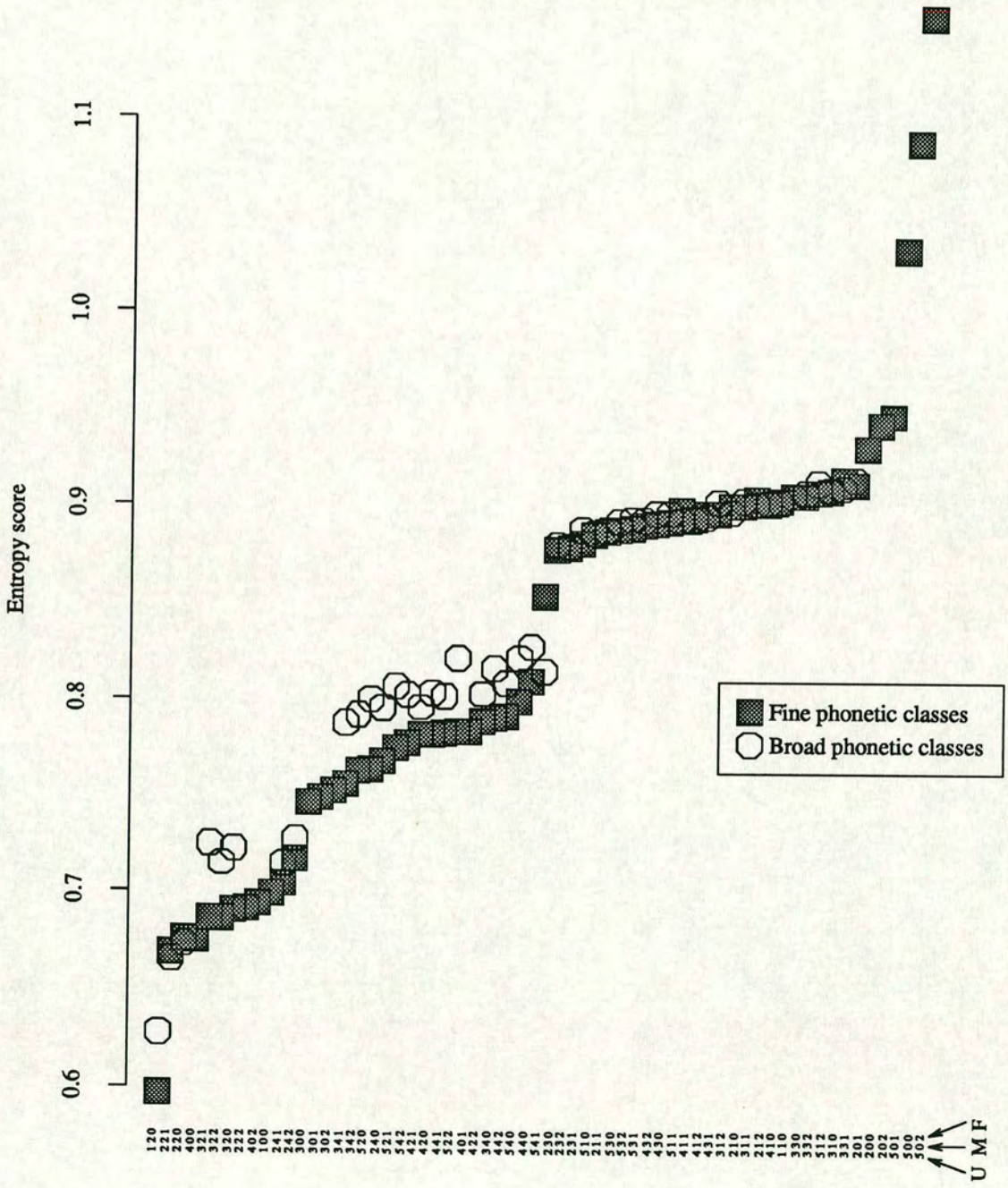


Figure 5.6: Energy feature evaluation: Entropy score ($P_{0,1}$)

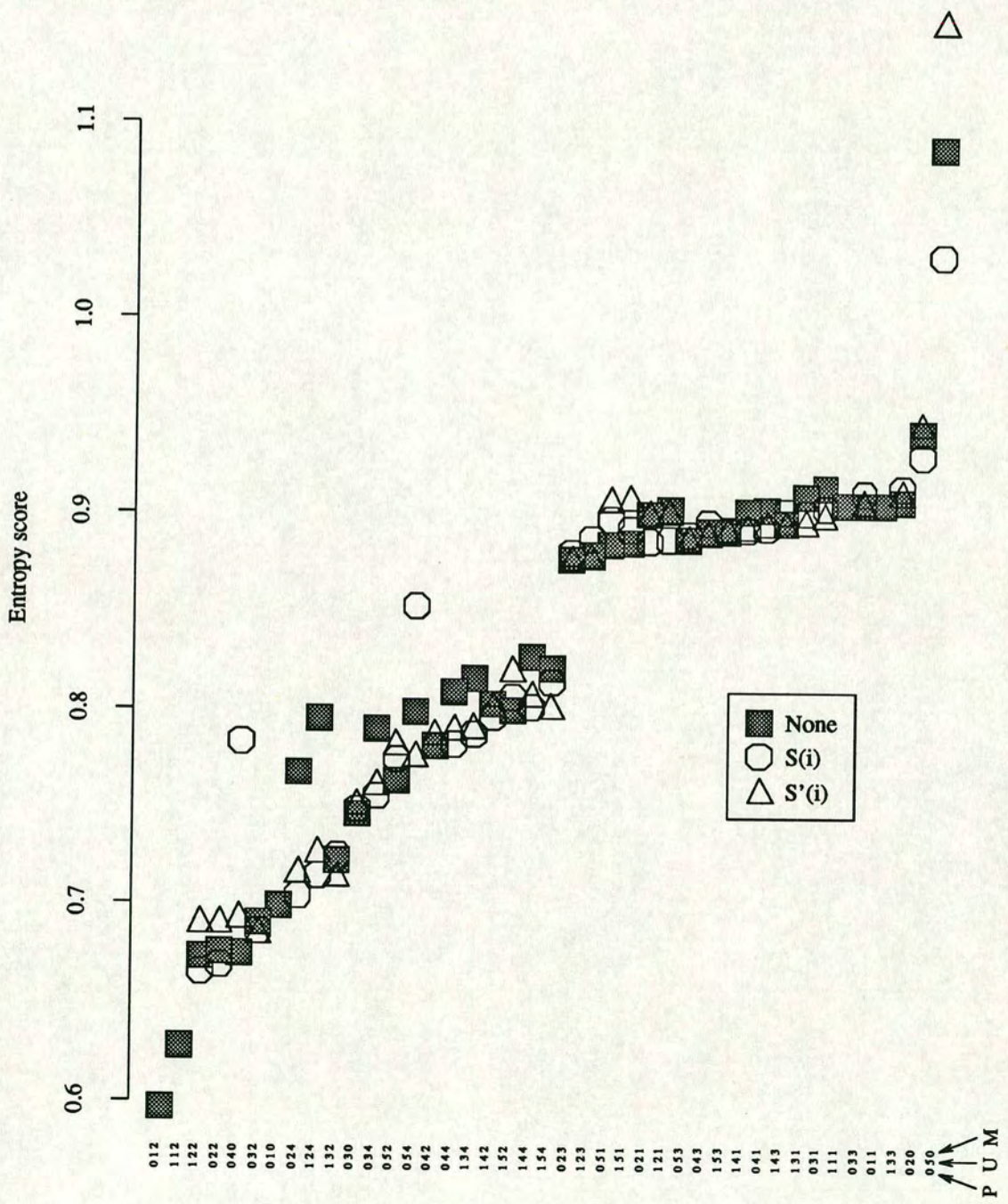


Figure 5.7: Energy feature evaluation: Entropy score ($F_{0,1,2}$)

energy and duration measures. An improvement in performance is generally obtained by using the maximum $z_{percentile}$ low-band energy measure M_2 rather than using the sum $z_{percentile}$ low-band energy measure M_4 of phones in a unit. Therefore this also provides evidence that sentential stress is correlated with an energy feature based on the peak of sonority in a syllable.

There is a significant difference in the entropy scores shown in Figure 5.6 between most $z_{percentile}$ based normalisations with statistics trained on fine phonetic classes and $z_{percentile}$ based normalisations trained on broad phonetic classes. Relatively little difference in the entropy scores is noticeable between the z_{mean} based normalisations trained on fine phonetic classes and broad phonetic classes. Since the optimal energy feature uses a $z_{percentile}$-transform and since the use of fine phonetic classes yields a lower entropy score, it is preferable to normalise for phone-types on a fine phonetic basis rather than on a broad phonetic basis. Two possible conclusions emerge from this with respect to the suitability of the training data — either the data provides a sample of the universal set of phone-types classified on a fine phonetic basis which represents the distribution of energy in such phone-types; or the broad phonetic classes shown in Table 5.1 form suboptimal groups.

Figure 5.7 reveals that normalisation for the number of phones in a unit has little effect on the performance of the Bayesian classifier for a given phonetic unit and low-band energy measure. Given that the optimal unit-type for the energy measure is the nucleus, no normalisation technique for the number of phones in a unit is applicable.

A final observation to make is in the comparison of the entropy score results for the duration and energy features. It is clear that the relative importance of duration and energy as cues to syllable prominence is dependent upon the choice of duration and energy features. No conclusions can therefore be drawn about the relative importance of these features in human perception of stress. Conclusions can only be drawn with respect to relative importance of the acoustic features investigated here for computer analysis of stress.

5.4 Optimisation of a relative energy feature

The peak-picking technique proposed in Section 4.5 is used to classify each syllable as either non-prominent or prominent depending on the magnitude of the low-band energy feature relative to its nearest neighbours. A syllable is classified as prominent if its low-band energy feature exceeds that of both its neighbours (end-points being inherently lower) and if such a local maximum in the low-band energy contour has a magnitude greater than some specified threshold K_{energy}. Syllables satisfying this criterion are labelled "*se*". A syllable is also classified as prominent if it corresponds with the maximum value of the low-band energy feature in the contour (independently of the threshold). Syllables corresponding to such maxima are labelled "*Se*". All other syllables are labelled "*ue*".

The low-band energy feature used corresponds to the syllable nucleus U_1 with $z_{percentile}$ low-band energy measures M_2 trained on fine phonetic classes P_0. Theoretically, this low-band energy feature is positive for prominent syllables and negative for non-prominent syllables. K_{energy} should therefore be optimal at 0.0. A plot of K_{energy} against the percentage of syllables which are assigned to the correct prominence category by the peak-picking technique (in a closed test) is shown in Figure 5.8. The optimal value of K_{energy} is -0.75.

There are a number of potential reasons as to why the optimal value of K_{energy} is not as expected. One possible reason is that the assumption made by the $z_{percentile}$ normalisation technique (that all of the lower low-band energy measures correspond to non-prominent syllables and that all of the higher low-band energy measures correspond to prominent syllables) is not valid. There are a number of prominent syllables for which the $z_{percentile}$ low-band energy measure is negative, and there are a number of non-prominent syllables for which the $z_{percentile}$ low-band energy measure is positive. By definition, there are the same number of prominent syllables with a negative $z_{percentile}$ low-band energy measure as there are non-prominent syllables with a positive $z_{percentile}$ low-band energy measure. It is empirically determined that reducing the threshold K_{energy} increases the number of non-prominent syllables with a $z_{percentile}$ low-band energy measure greater than the threshold more than it decreases the number of prominent syllables with a $z_{percentile}$ low-band energy measure less than the threshold.

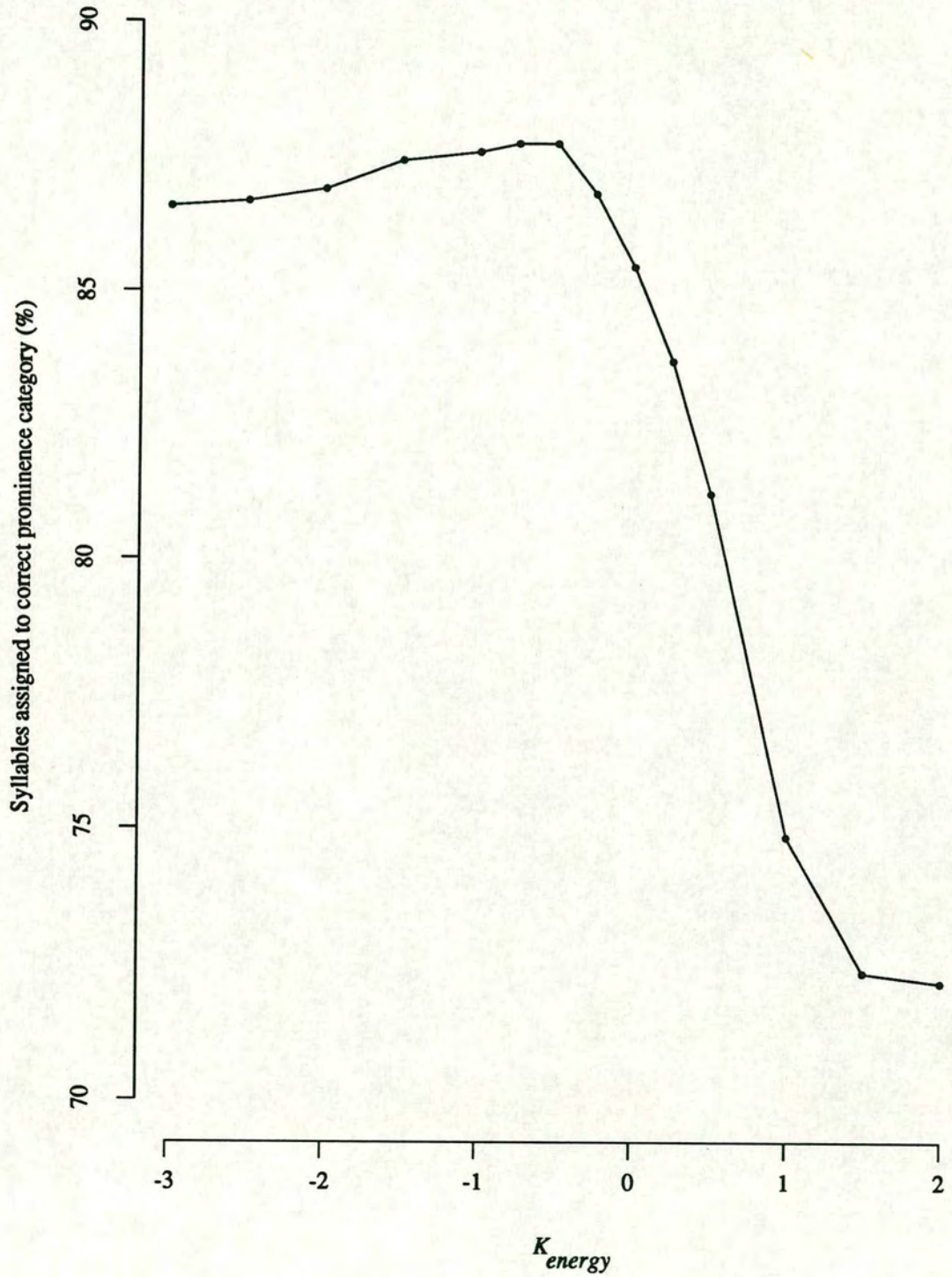


Figure 5.8: Optimisation of K_{energy} threshold

If it is assumed that a syllable is categorised as prominent only if its corresponding $z_{percentile}$ low-band energy measure is greater than K_{energy}, then a greater number of syllables should be incorrectly categorised (as either prominent or non-prominent) by reducing K_{energy}. This is contrary to the evidence provided in Figure 5.8. It is therefore unlikely that the optimal value of K_{energy} is less than expected because the assumption made by the $z_{percentile}$ normalisation technique is not strictly valid.

The above argument assumes that a syllable is categorised as prominent only if its corresponding $z_{percentile}$ low-band energy measure is greater than K_{energy}. The peak-picking technique, however, enforces the additional constraint that the syllable must correspond to a local maximum in the low-band energy contour in order to be categorised as prominent. This is a consequential factor in the optimal value of K_{energy} being lower than expected. There are more prominent syllables (22.3%) in the training data which correspond to a local maximum in the low-band energy contour than non-prominent syllables (8.0%). Optimisation of the threshold K_{energy} minimises the number of these (22.3% + 8.0%) syllables which are assigned to the incorrect prominence category. Reducing K_{energy} from 0.0 to -0.75 increases the number of prominent syllables labelled as "*se*" more than it decreases the number of non-prominent syllables labelled as "*ue*". In other words, the optimisation of K_{energy} is dependent upon the distributions of the $z_{percentile}$ low-band energy measure which correspond to local maximum in the energy contour for prominent and non-prominent syllables.

The optimisation of the threshold K_{energy} is also dependent upon the error measure used to evaluate the performance of the peak-picking technique. Here, the total-error rate is used because it is equally important to classify a given syllable correctly as non-prominent as it is to classify a given syllable correctly as prominent. If the classification criterion was to ensure that the same percentage of prominent syllables are correctly classified as non-prominent syllables, then an equal-error rate would be a more appropriate measure. A more robust measure (such as the entropy score used for the Bayesian classifier) for the performance of the peak-picking technique would produce a different optimal value for K_{energy}.

An entropy score cannot be produced by the peak-picking method. Its performance is therefore compared with the performance of the Bayesian classifier by examining the

		Peak-picking algorithm label			
		Se	*se*	*ue*	total
Hand Label	*P*	422 (7.9%)	1101 (20.7%)	575 (10.8%)	2098 (39.4%)
	NP	38 (0.7%)	236 (4.4%)	2947 (55.4%)	3221 (60.6%)
	total	460 (8.6%)	1337 (25.1%)	3522 (66.2%)	5319 (100.0%)

Correct classification rate = 4470/5319 (84.0%)

Table 5.2: Energy: Confusion matrix for peak-picking algorithm
(*Se* — syllable with maximum energy of contour; *se* — syllable corresponding to a local peak in the energy contour greater than K_{energy}; *ue* — unstressed syllable on the basis of energy; *P* — prominent syllable; *NP* — non-prominent syllable)

		Bayesian classifier label		
		P	*NP*	total
Hand Label	*P*	1650 (31.0%)	448 (8.4%)	2098 (39.4%)
	NP	516 (9.7%)	2705 (50.9%)	3221 (60.6%)
	total	2166 (40.7%)	3153 (59.3%)	5319 (100.0%)

Correct classification rate = 4355/5319 (81.9%)

Table 5.3: Energy: Confusion matrix for Bayesian classification
(*P* — prominent syllable; *NP* — non-prominent syllable)

percentage of syllables which are assigned to the correct prominence category in an open test. The levels of agreement between the prominence category of a syllable as transcribed by hand and as assigned automatically are shown in Tables 5.2 & 5.3 for the peak-picking algorithm and for the Bayesian classifier respectively. As well as giving an improvement of syllable prominence classification, the simple peak-picking algorithm has the additional benefits over the Bayesian classifier in requiring less computation (and hence it is faster) and not requiring a training stage.

5.5 Summary

The investigation described above aimed to determine a low-band energy feature whose distributions for prominent and non-prominent syllables have the greatest separability. The entropy score obtained by a Bayesian classifier was used as a robust measure of this separability for 113 different low-band energy features. The optimal low-band energy feature $E_{feature}$ of those investigated is given by:

$$E_{feature} = \frac{E_{nucleus} - \varphi_E^p(fine\_phone\_class)}{\sigma_E(fine\_phone\_class)} \tag{5.3}$$

where $E_{nucleus}$ is the low-band energy of a syllable nucleus, $\varphi_E^p(fine\_phone\_class)$ and $\sigma_E(fine\_phone\_class)$ are the p'th percentile and the standard deviation of phone-level low-band energies for phones classed on a fine phonetic basis respectively, and p is the percentage of non-prominent phones in the training data.

A simple peak-picking algorithm applied to this low-band energy feature is shown to give an improvement of syllable prominence classification over the Bayesian classifier, while requiring less computation and not requiring an additional training stage. In an open test, 84.0% of syllables are given the same prominence status by the peak-picking algorithm as when transcribed by hand.

The transcription (by hand) of the perceived prominence of each syllable in the training and test data cannot be regarded as definitive because there are instances when the discrete categorisation of a syllable as prominent or non-prominent is ambiguous. There is therefore some degree of error in the comparisons of prominence levels labelled by hand and labelled by the automatic procedures discussed above. The level of agreement between the prominence label assigned by the peak-picking algorithm and as transcribed by hand, is therefore regarded as a comparative measure only. By comparison to the duration feature $D_{feature}$ (Equation 4.6) with a 77.8% agreement level, it can be concluded that the low-band energy feature $E_{feature}$ (Equation 5.3) has a higher correlation with sentential stress for the speech of the one native English talker of the training and test data. It is important to emphasis that it cannot be concluded that the energy feature $E_{feature}$ necessarily has a higher correlation with sentential stress than the duration feature $D_{feature}$ for all speakers, or that energy has a greater importance than duration as

a cue in the human perception of sentential stress.

Chapter 6

Vowel quality measures

In addition to duration and energy features, it is suggested in Section 3.1 that perceived vowel quality is also correlated with sentential stress. A number of measures of vowel quality are proposed in this Chapter. Vowel quality measures are based on the assumption that the nuclei of prominent syllables are well articulated and are less affected by contextual assimilation, relative to the nuclei of non-prominent syllables. The aim here is to devise a measure of vowel quality based on formant frequency trajectories which is correlated with the prominence of a syllable.

The trajectories of formant frequencies through each vowel of an utterance are estimated using a formant tracking algorithm based on generalised centroids (Crowe & Jack, 1987; Crowe, 1988). The speech waveform, sampled at 20kHz, is Fourier-transformed every 5ms using a 512-point FFT with a 20ms Hann window. This ensures that formant frequency values are synchronised with the energy contour and *FØ* contour calculations. High-frequency pre-emphasis is applied to the resultant spectrum, and a triple centroid is estimated over the frequency range from 234Hz (the 6'th frequency bin) to 2891Hz (the 74'th frequency bin). The formants *F1*, *F2*, and *F3* are captured within this range while *FØ* and higher formants are excluded.

The formant contours generated by the generalised centroids algorithm are post-processed by a five frame median filter and three frame Hann window to remove isolated spurious formant frequency estimates (Section 7.3.1).

6.1 Vowel context and speaker normalisation

The perceptual quality of a vowel is correlated with the trajectories of the formant frequencies *F1*, *F2*, and *F3*. However, these trajectories vary depending on phonetic context and the individual speaker (Peterson & Barney, 1952). Given that there are physical constraints on the articulators, two different vowels pronounced in different phonetic contexts by the same speaker may have similar formant frequency values. Similarly, a vowel perceived as similar when spoken in different contexts by the same speaker may have different formant trajectories (phonetic context dependency). The acoustic realisation of a vowel described by the same phonetic symbol and spoken by two speakers may differ because of anatomical and physiological differences between the speakers. This may result in different vowels pronounced by different speakers having similar formant trajectories, and a vowel perceived as similar when spoken by two different speakers having different formant trajectories (speaker dependency).

A number of normalisation techniques are investigated here which aim to compensate for these context-dependent and speaker-dependent variations in the formant trajectories. In previous investigations of vowel normalisation (Disner, 1980; Syrdal & Gopal, 1986; Hillenbrand & Gayvert, 1993) the focus is placed on minimising the variance within a set of vowels spoken by different speakers which are presumed to correspond to the same vowel target, while maximising the separation of sets of vowels which are presumed to correspond to different vowel targets. Here, the focus of the vowel normalisation is to maximise, for each set of vowels presumed to correspond to the same vowel target, the separation of vowels which form the nuclei of prominent and non-prominent syllables.

The phonetic context dependency of the formant frequency trajectories is compensated for by obtaining one measurement from each trajectory within a vowel. Three different trajectory features $C_{1,2,3}$ are investigated.

- C_1, the peak/valley/mid-point value of a frame-level formant trajectory. The peak/valley/mid-point value of a formant trajectory over a vowel is determined using the method proposed in Section 5.1.

- C_2, the mid-region mean value of a formant trajectory. The mid-region mean value is defined as the average formant frequency value over the central half of a vowel.

Formant frequency values in the initial and final quarters of the vowel are assumed to be highly influenced by segmental context and are disregarded.

- C_3, the *stable*-region mean value of a formant trajectory. The stable-region of a vowel is defined after van Bergem (1988). The *F1*, *F2*, and *F3* contours are initially transformed to a Mel-scale using Equation 6.1.

$$F_{Mel} = 2595.0 \log\left(1.0 + \frac{F_{Hz}}{700.0}\right) \tag{6.1}$$

A window of quarter of the vowel length is moved through the *F1*, *F2*, and *F3* formant trajectories in search of a region of the vowel for which the pooled variance of these (Mel-scaled) formants is a minimum. The average formant frequency values (in Hertz) are determined within this region. If the duration of the vowel is so short that the window fails to occupy at least two frame-level formant frequency values (and hence one is unable to determine a variance) then a *break* value of infinity is associated with the vowel. Since the frame-level formant frequency values are specified at intervals of 5ms, vowels shorter than 40ms are associated with the break value.

Syrdal & Gopal (1986) propose that the speaker dependency of the formant frequency values can be normalised by use of formant differences on the critical-band-based Bark scale, $F2_{Bark} - F1_{Bark}$, $F3_{Bark} - F2_{Bark}$. The set of formant frequency measurements for each vowel are converted to a Bark-scale using Equation 6.2 (Zwicker & Terhardt, 1980).

$$F_{Bark} = 13.0 \arctan(0.76 F_{kHz}) + 3.5 \arctan\left(\frac{F_{kHz}}{7.5}\right)^2 \tag{6.2}$$

A problem in using formant differences, however, is that some pairs of phonetically distinct vowels have similar formant differences — for example, the pairs /ɑ, ɔ/ and /u, ʊ/ (Hillenbrand & Gayvert, 1993). Syrdal & Gopal address this problem by including the difference between *F1* and *FØ* ($F1_{Bark} - F\emptyset_{Bark}$) as a parameter in the normalisation.

In summary, each vowel in an utterance is characterised by a set of measurements derived from *FØ* and formant frequency trajectories. One value is obtained from each frequency trajectory within a vowel by using one of three methods — $C_{1,2,3}$. For ex-

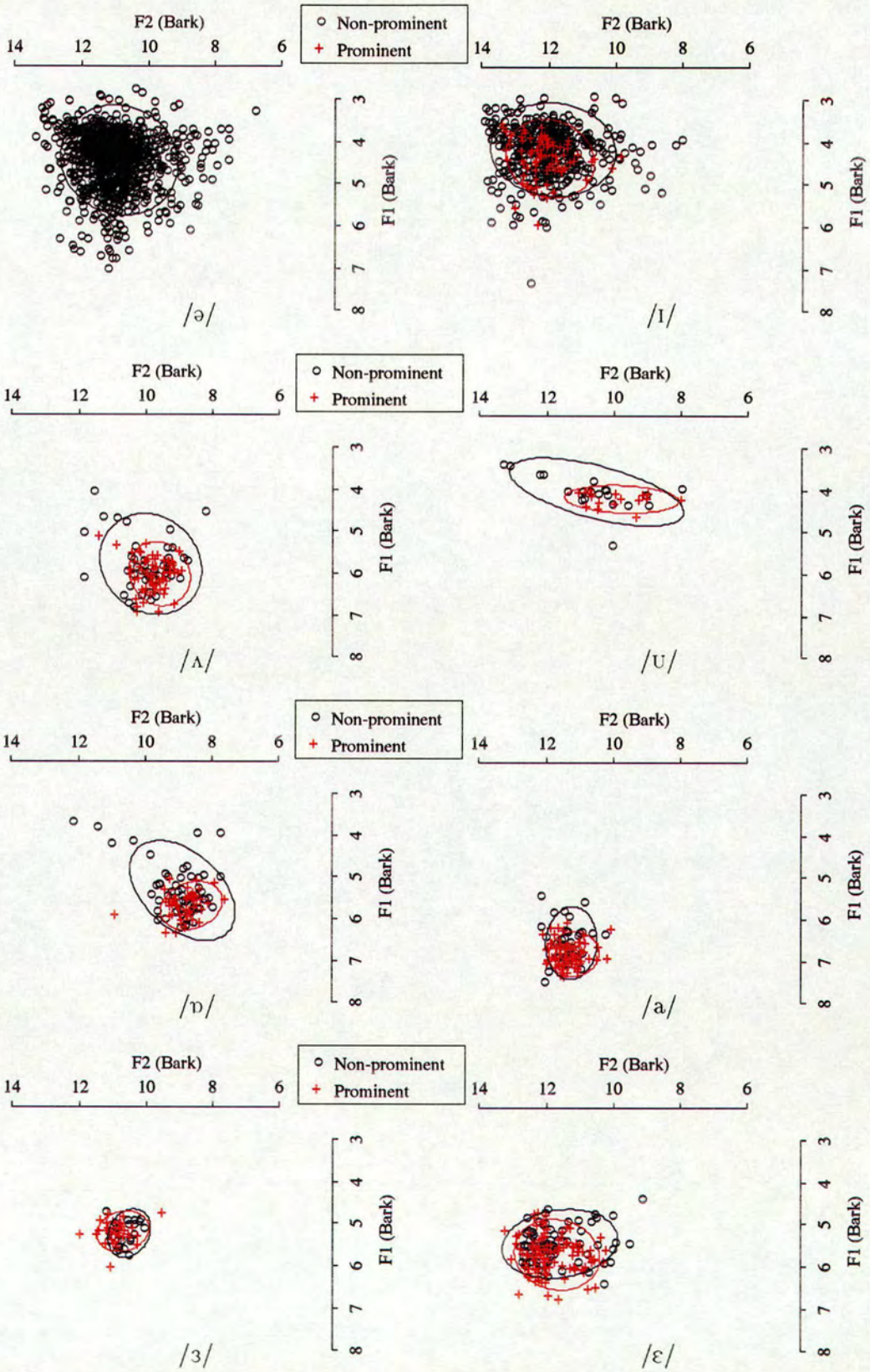
ample, four values can be obtained from the *FØ*, *F1*, *F2*, and *F3* trajectories within a vowel by calculating the mid-region mean value C_2 for each trajectory. A vowel can be characterised by these values converted to a Bark-scale or by the difference between them.

6.2 Vowel target model

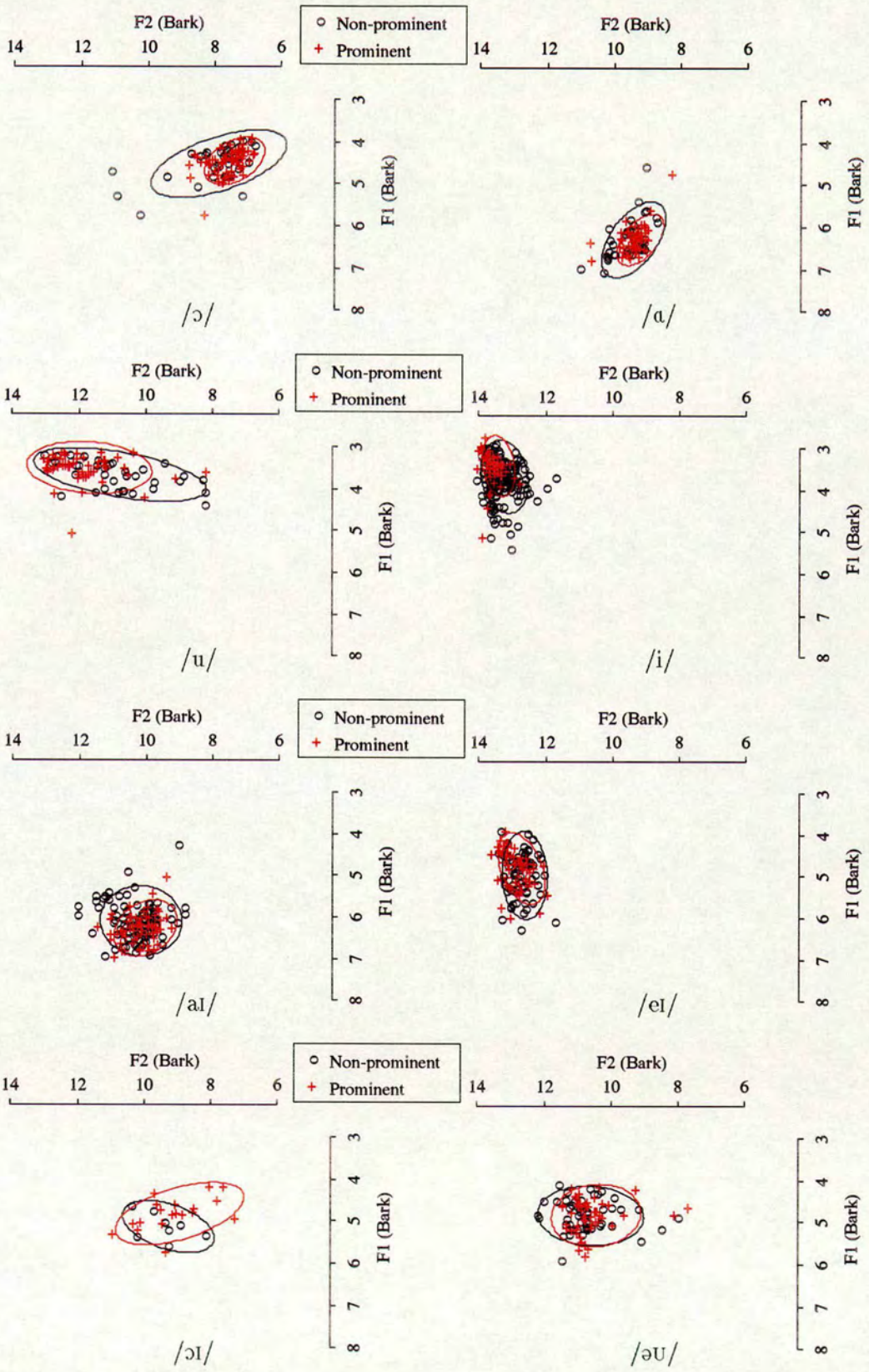
Figure 6.1 shows the clustering of prominent and non-prominent vowels for the twenty vowel-types in the training data of English speech. The ellipses indicate the degree of spread in the clusters of vowels in prominent syllables (shown in red) and in non-prominent syllables (shown in black). A 2-dimensional probability density function (p.d.f.) is associated with the variables plotted along the axes (in this case, the Bark-scaled mid-region mean values C_2 of *F1* and *F2*). The intersection of this p.d.f. (which can be imagined as a 'hill') and a plane at a level of two times the standard deviation (a slice through the hill) forms an ellipse when projected onto the zero-zero plane of the p.d.f. (the floor of the hill). It is these ellipses which are shown in Figure 6.1. There is a tendency for the vowels (at least the monophthongs) in prominent syllables to cluster more tightly than corresponding vowels in non-prominent syllables, and a tendency for the clusters of vowels in prominent syllables to lie within the clusters of the corresponding vowels in non-prominent syllables. This is not necessarily applicable to diphthongs because the inherent glide in their formant trajectories is not modelled by the single measurement associated with each formant trajectory.

The populations of formant frequency measurements approximate a Normal (Gauss-Laplacian) distribution in the Bark domain. The application of the Bayesian classifier to model each type of vowel using Bark-scaled frequency measurements is therefore viable[1]. A Bayesian model is created for each vowel-type. Each vowel is characterised by a number of features. The features are a combination of the absolute or the differences in Bark-scaled formant frequencies and fundamental frequency, where each of the formant frequencies and the *FØ* value associated with the vowel are obtained using any one of

[1]The requirement for the input data to be Normally distributed is generally not a prerequisite of a Bayesian classifier. This constraint is imposed here by the implementation of the classifier as described in Appendix B



continued on next page...



continued on next page...

... continued from previous page

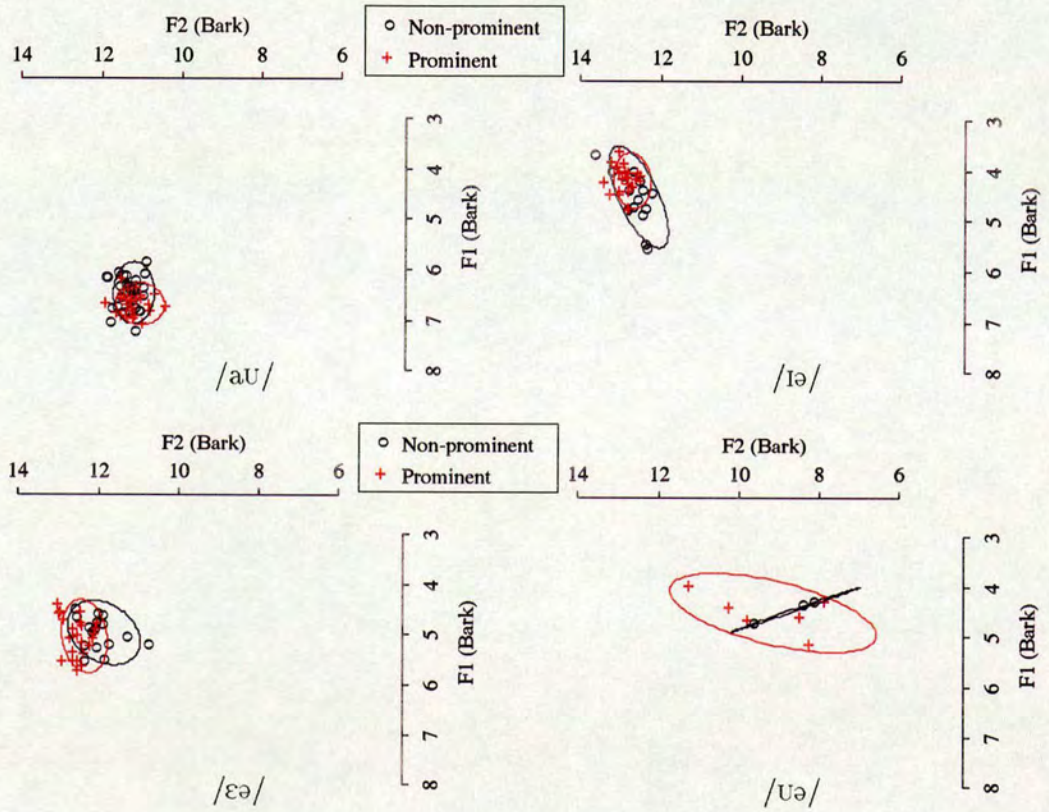


Figure 6.1: Vowel targets in prominent and non-prominent syllables

the trajectory features $C_{1,2,3}$. The vowel target models are calculated from frequency measurements of every vowel token (those corresponding to the nuclei of both prominent and non-prominent syllables) in the training data.

The quadratic discriminant score Q_{vowel} is determined for each vowel token in the training data with respect to its corresponding vowel target model. This score is a weighted variance-normalised distance measure between the vector of features which characterise the vowel and the centroid of the respective vowel target model. Thus, Q_{vowel} is a measure of how close the vowel is to its target.

It is assumed that vowels closer to their respective vowel target model correspond to the nuclei of prominent syllables. Under this assumption, Q_{vowel} for a vowel in a prominent syllable is less than Q_{vowel} for a vowel in a non-prominent syllable. Thus, in a

distribution of Q_{vowel} for a given vowel-type containing p percent of vowels corresponding to non-prominent syllables, the lower $(100 - p)$ percent of Q_{vowel} values are assumed to correspond to vowels in prominent syllables. The distance of a vowel from its respective vowel target model is expressed relative to the $(100 - p)$'th-percentile of the quadratic discriminant scores for its vowel-type $\varphi_Q^{(100-p)}(vowel\_type)$.

$$Q' = Q_{vowel} - \varphi_Q^{(100-p)}(vowel\_type) \tag{6.3}$$

If the assumption is correct then Q' is positive for all vowels in non-prominent syllables and is negative for all others. The population of Q' does not have a Normal distribution. Since the implementation of the Bayesian classifier (described in Appendix B) assumes that the input has a Normal distribution, it is not used to classify the prominence of a syllable on the basis of Q'.

A simple peak-picking algorithm is used to label the contribution made by vowel quality to the perceived prominence of a syllable. A syllable is labelled as "*sq*" to indicated that it is prominent relative to its neighbours on the basis of vowel quality; that is, if Q' for the vocalic nucleus of the syllable is less than zero and if Q' for the nucleus of the syllable is less than Q' for the nuclei of both neighbouring syllables (end-points being inherently higher). Otherwise, a syllable is labelled as "*uq*", indicating that is not prominent relative to its neighbours on the basis of vowel quality. Schwa never forms the nucleus of a prominent syllable. Hence, Q' is positive for all syllables with a schwa nucleus, by definition. They are therefore labelled "*uq*". Q' is set to 'infinity' for any syllables which have a syllabic consonant as the nucleus. They are, therefore, also labelled "*uq*".

Every syllable in the test data is labelled using the above criteria. These labels are compared with prominence levels assigned to syllables by hand and the level of agreement between them is determined. Agreement levels are represented in Figure 6.2 for when the vowel target models are calculated from a number of different combinations of features. The highest level of agreement (71.6%) is obtained when the vowel target models are calculated from Bark-scaled peak/valley/mid-point (C_1) frequency differences $F1_{Bark} - F\emptyset_{Bark}$, $F2_{Bark} - F1_{Bark}$, and $F3_{Bark} - F2_{Bark}$. A confusion matrix for the prominence labels assigned using this method versus prominence labels assigned by hand

		Vowel target label		
		sq	*uq*	total
Hand Label	*P*	963 (18.1%)	1135 (21.3%)	2098 (39.4%)
	NP	375 (7.1%)	2846 (53.5%)	3221 (60.6%)
	total	1338 (25.2%)	3981 (74.8%)	5319 (100.0%)

Correct classification rate = 3809/5319 (71.6%)

Table 6.1: Vowel quality: Confusion matrix for peak-picking algorithm
(*sq* — syllable corresponding to a local peak in the vowel quality contour less than 0.0;
uq — unstressed syllable on the basis of vowel quality; *P* — prominent syllable;
NP — non-prominent syllable)

is shown in Table 6.1.

In order to place the agreement levels shown in Figure 6.2 into perspective, consider the performance of the peak-picking algorithm given the $z_{percentile}$ measure of a random variable with a Normal distribution. The thresholds for each vowel-type are given by the $(100 - p)$'th-percentile of a Normal distribution, where p represents the percentage of vowels which are non-prominent in the training data. Every syllable in the test data is associated with a random number from a single Normal distribution. A syllable is labelled as prominent if this random number is less than the threshold associated with the vowel-type of the nucleus of the syllable. Approximately $69 \pm 2\%$ of the syllable labels determined by this method are the same as those given by hand.

If the features used to characterise a vowel include an absolute measure of fundamental frequency (cf. $F1_{Bark} - F\emptyset_{Bark}$) or are based on stable-region mean values (C_3) then the correlation of the vowel target distance Q' with sentential stress is worse than the correlation of a $z_{percentile}$-transformed, Normally-distributed random variable with stress.

A measure of the distance from a vowel target can, in some cases, be more effective at distinguishing between prominent and non-prominent syllables than a random variable in the analysis of the speech of a native speaker of English. However, such a measure cannot be used in the analysis of the speech of a non-native speaker because of vowel pronunciation errors.

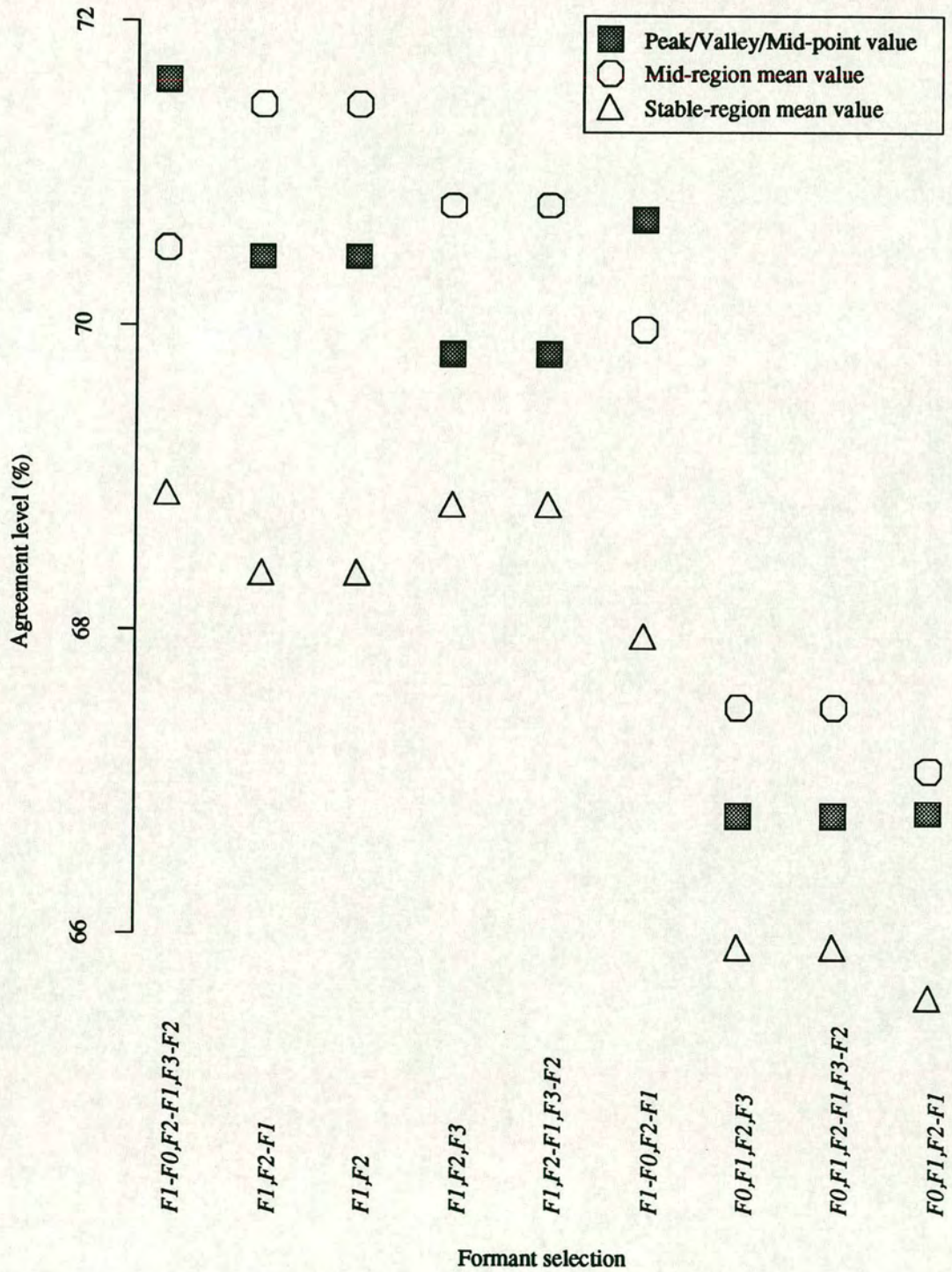


Figure 6.2: Vowel quality feature evaluation

There are two types of vowel pronunciation errors for non-native speakers of a language. Firstly, vowels which are assigned the same phonetic symbol in the target language to be acquired by a learner and in the native language of the learner, can differ in their phonetic realisation (Disner, 1980). Thus, vowel quality in the target language is influenced by the native language of the learner, as well as by context (both segmental context and syllable prominence). Secondly, vowels which do not belong to the vowel system of the learner's native language tend to be substituted by a vowel in the learner's native vowel system which is perceived (in the mind of the learner) to be 'close' to the desired vowel in the target language (di Benedetto *et al.*, 1992).

There is a fundamental problem in using a measure of the quality of a vowel spoken by a foreign language learner as a correlate of the perceived prominence of a syllable. A measure is required to determine if the quality of the vowel spoken by the learner is comparable to that of a native speaker when pronounced in the same context. This may be done by comparing the acoustic features which characterise the learner's vowel with the vowel system of a native speaker. The problem is that a mismatch between the vowel token pronounced by the learner and the intended vowel target for the target language can be due to a number of different reasons, other than the perceived prominence of a syllable. A mismatch can be due to anatomical and physiological difference between the learner and a native speaker, due to differences in the phonetic realisation of the vowel in the source and target languages, or due to a pronunciation error by the learner. This problem can be avoided only by comparing the acoustic features which characterise the learner's vowel with a vowel target modelled in the learner's vowel system (rather than in the vowel system of a native speaker of the target language). Therefore, if vowel quality is to be used in the automatic analysis of stress in utterances spoken by a non-native speaker, the task of learning correct vowel pronunciation and the task of learning the prosodic aspects of the target language must be kept separate.

The results shown in Figure 6.2 indicate that a measure of the distance of a vowel from its respective target is comparable to a $z_{percentile}$-transformed, Normally-distributed random variable when used as a correlate of the perceived prominence of a syllable. In the investigation, a vowel target is modelled from vowel tokens of a native speaker of English. The vowels used in the test data are spoken by the same speaker. In is inferred

that even if the tasks of learning correct vowel pronunciation and learning the prosodic aspects of the target language are kept separate and a non-native speaker has successfully learnt to approximate the vowel system of the target language, then it is not possible to use the vowel-to-vowel-target distance as a correlate of stress.

6.3 Vowel stability measure

It is assumed here that a continuum exists in perceptual vowel quality from when a vowel is enunciated in isolation, to when a vowel is pronounced as the nucleus of a prominent syllable, to when a vowel is pronounced as the nucleus of a non-prominent syllable, to a schwa. Van Bergem (1994) proposes that schwa is, "a vowel without articulatory target that is competely assimilated with its [segmental] context" and that reductions in perceptual vowel quality (such as that which can occur in the nuclei of non-prominent syllables) is the partial assimilation of vowels with their segmental context. The formant trajectories of schwa are therefore completely dependent upon the position of the articulators during the production of surrounding phones, and vowels which have a perceptual reduction in quality are subject to some degree of contextual assimilation. The hypothesis that formant trajectories (or the spectra) are more changeable in the vocalic nuclei of non-prominent syllables (which are subject to vowel reduction) than in the vocalic nuclei of prominent syllables, is considered here.

Seven different measures of the degree of change in formant trajectories and one measure of the degree of spectral change through the course of a vowel, are investigated.

- V_1. The peak/valley/mid-point values in the *F1* and *F2* trajectories are determined using the method described in Section 5.1. An accumulative score is associated with each trajectory based on the type of point which is located — peak (+3), valley (+2), and rise or fall (+1). Thus, for example, if the trajectories of both *F1* and *F2* contain local peaks then a score of +6 is assigned to the syllable; and if the *F1* trajectory contains a local peak and the *F2* trajectory contains a fall then a score of +3 is assigned to the syllable.

- V_2. Maximum pooled variance of $F1_{Mel}$, $F2_{Mel}$, and $F3_{Mel}$ is calculated over the duration of the vowel using a window of quarter of its length — an 'instability'

measure. This measure approximates a log-Normal distribution.

- V_3. Minimum pooled variance of $F1_{Mel}$, $F2_{Mel}$, and $F3_{Mel}$ is calculated over the duration of the vowel using a window of quarter of its length — a 'stability' measure. This measure approximates a log-Normal distribution.

- $V_{4,5,6,7}$. Maximum V_4, minimum V_5, mean V_6 or standard deviation V_7 (four different measures) is determined for the rate of change of $F1$, $F2$, and $F3$ from one frame to the next, across the entire duration of the vowel.

- V_8. The average Euclidean distance of frame-to-frame log-power spectra is calculated over the central 50%[2] of syllable nuclei — a measure of 'spectral change'. This measure is determined by performing a Fourier transform to frames (20ms in duration) of speech data at regular intervals (every 5ms) in an utterance and calculating a log-power spectrum for each frame. The Euclidean distance (Equation 6.4) is evaluated for adjacent spectra. This distance is averaged over the central 50% of each syllable nucleus in the utterance.

$$Euclidean_{ij} = \sqrt{\sum_{n=1}^{N}(x_{in} - x_{jn})^2} \tag{6.4}$$

where N is the number of frequency bins in a spectrum and x_{in} is the log-power in the n'th frequency bin of the spectrum of the i'th frame ($j = i + 1$).

A syllable is labelled as prominent on the basis of vowel stability if V_1 equals +6; otherwise it is labelled as non-prominent. Each of the measures $V_{2,\ldots,8}$ is used as a unidimensional feature for the Bayesian classification of syllable prominence. None of the measures succeed in classifying syllable prominence better than a $z_{percentile}$-transformed, Normally-distributed random variable. The use of any measure results in more prominent syllables being labelled automatically as non-prominent than being correctly labelled as prominent. The best performance (68.8% of syllables in agreement with the prominent labels transcribed by hand in the test data) is yielded by the measurement V_8.

[2]The central 50% of a syllable refers to the mid-portion of the syllable, thus excluding 25% at the beginning of the syllable and 25% at the end.

6.4 Phonological rules

According to the phonological description of English R.P., none of the lax vowels /ɪ, ɛ, ʌ, ʊ, ɒ, (æ)/ can appear in prominent open syllables (Ladefoged, 1982). It may seem reasonable to use a rule that labels all lax vowels that occur in open syllables as non-prominent. There are, however, two disadvantages to the use of such a rule. Firstly, the rule is dependent upon the definition of a syllable. An open syllable defined phonologically using the principle of maximal onset (Pulgram, 1970) will differ from one defined using the principle of maximal coda or one defined using acoustic-phonetic criteria (Section 8.1.1). Secondly, foreign speakers may stress any of the vowels /ɪ, ɛ, ʌ, ʊ, ɒ, (æ)/ in open syllables when speaking English if such syllable structures are permissible in their mother-tongue. In the training data, spoken by a native speaker of British English, the only vowels which form the nucleus of prominent open syllables in word final position are long, tense vowels and diphthongs /ɑ, i, ɔ, u, aɪ, aɪ, eɪ, ɔɪ, əʊ, ɛə, ɪə, ʊə/. This does not mean that every speaker of English (including learners of English with strong foreign accents) only place stress on word-final open syllables containing this set of phones. Recall that the aim here is to identify the location of prominent syllables so that a foreign speaker can be informed about incorrect stress placement in English, such as placing stress on lax vowels in word-final open syllables. It is therefore inappropriate to assign all word-final open syllables with lax vowel nuclei as non-prominent by default.

6.5 Summary

The vowel quality of a syllable nucleus is assumed to be an acoustic correlate of stress and a number of measures of vowel quality based on formant trajectories are proposed in this Chapter.

A normalised quadratic discriminant score (Equation 6.3) is proposed as a measure of how close a vowel is to its target, where a vowel target is a statistically trained model based on frequency measurements which aim to be normalised for phonetic context and speaker dependencies. This measure is used to label syllables spoken by a native English speaker as either prominent or non-prominent and yields a 71.6% agreement level with prominence labels based on human perception. It is argued that this measure cannot

be used as a correlate of stress in utterances spoken by non-native speakers of English because of vowel pronunciation errors.

A number of different measures of the degree of change in formant trajectories (and spectral change) due to contextual assimilation are investigated. The average Euclidean distance of frame-to-frame log-power spectra (Equation 6.4) is used as a measure of the degree of spectral change in a vowel which is due to contextual assimilation. This measure yields a 68.8% level of agreement between prominence labels assigned automatically and as transcribed by hand.

The underlying assumptions made are that, in connected speech, the vocalic nuclei of prominent syllables are closer to their respective vowel targets than the vocalic nuclei of non-prominent syllables, and that formant trajectories are more changeable in reduced vowels than fully articulated vowels because they are subject to greater degrees of contextual assimilation. The proposed measures of vowel quality fail to convincingly support these assumptions since their use to distinguish prominent syllables from non-prominent syllables is comparable with using of a $z_{percentile}$-transformed, Normally-distributed random variable as an input parameter to a peak-picking algorithm. The application of the proposed vowel quality measures as correlates of stress is therefore inappropriate. The use of phonological rules to transcribe syllables as non-prominent under certain conditions, is also dismissed.

The formant trajectories of vowels are known to vary within the vowel system of a speaker. The cause of these variations are not yet fully understood although it has been suggested that segmental context and stress play a role. Duration and articulatory constraints may, however, be the means whereby segmental context and stress influence formant trajectories (Lindblom, 1963). Whether or not stress affects the perception of vowel quality directly, the prominence of a syllable cannot be determined from a measure of vowel quality, in the same way as it is not possible to determine the segmental context of a vowel from its perceived quality, without first understanding the exact nature of the influence of a vowel's context on its formant trajectories in connected speech. It is beyond the scope of this thesis to formulate such an understanding. Research to systematically measure coarticulatory effects on the formant frequency trajectories of the schwa in isolated nonsense words is reported by van Bergem (1994).

To conclude, Chapters 4 & 5 and this Chapter have concentrated on optimising the extraction of acoustic correlates required for the prosodic analysis of English. The underlying principle is to normalise the acoustic parameters for variations due to non-prosodic aspects of speech. The duration, energy and vowel quality features investigated are relevant to spoken English. Thus, the acoustic features presented might not be applicable to other languages. For example, the duration feature $D_{feature}$ (Equation 4.6) might not be related to the syllable lhyme for languages which differ from English in their syllabic structure, rhythm and timing. The methodology outlined in these Chapters can, however, be applied to investigations of the acoustic correlates of prosodic aspects of languages other than English.

Chapter 7

Fundamental frequency extraction

Methods to automatically determine the fundamental frequency of a speech waveform are addressed in this Chapter. The most reliable and accurate method of determining the fundamental frequency of a speech waveform is sought in order to minimise the number of errors occurring during *FØ* extraction from propagating into the prosodic analysis.

A preliminary evaluation of the FDAs reviewed in Section 3.2 provided evidence (which is supported by the evaluation reported in Section 7.2) that the super resolution *FØ* determinator (SRFD) and the feature-based *FØ* tracker (FBFT) perform most reliably and consistently for both male and female speech. A number of modifications to the SRFD algorithm (summarised in Section 3.2.6) are proposed in Section 7.1 in order to reduce the occurrence of errors such that it is optimised for prosodic analysis.

The performances of the *enhanced* SRFD algorithm and the *FØ* determination algorithms summarised in Section 3.2 are evaluated in Section 7.2. The evaluation uses laryngograph data to generate a reference contour and is tailored to identify the most appropriate method of fundamental frequency extraction of adult male and female speech recorded with little ambient interference.

A de-step filter is proposed in Section 7.3.2 to post-process the *FØ* contour generated by an FDA in a way which further reduces the occurrence of errors.

7.1 Enhanced super resolution *FØ* determinator (*e*SRFD)

The speech waveform is initially low-pass filtered to simplify the temporal structure of the sampled waveform. Each frame of filtered sample data is then processed by the 'silence' (low energy) detector described in Section 3.2.6.

The definitions of the segments x_n and y_n (Equation 3.8) are refined in order to improve the synchronisation of the *FØ* contour with the speech data. In the unmodified implementation of SRFD, the r'th frame of an utterance starts at $t_{interval}(r-1)$ milliseconds into the utterance, where $t_{interval}$ is the interval between successive analysis frames ($t_{interval}$ = 6.4ms for compatibility with the other FDAs during the evaluation and $t_{interval}$ = 5.0ms to synchronise *FØ* values with the energy contour during prosodic analysis). The analysis of each frame yields an estimated value of the fundamental period as n_0. In other words, the fundamental period for the two cycles following the start of the frame is n_0 samples. Thus, the *FØ* value produced for a frame describes the data centred $t_{interval}(r-1)+n_0/F_s$ milliseconds into the utterance (where F_s is the sampling frequency in kHz). This is dependent upon the estimated fundamental period of the analysis frame. An asynchronous *FØ* contour is therefore generated. To rectify this problem, the first sample of the segment y_n is defined as the sample $s(1)$, $t_{interval}(r-1)$ milliseconds into the utterance for the r'th frame. Each frame of speech data contains a set of samples $s_N = \{s(i) \mid i \in -N_{max}, \ldots, N - N_{max}\}$ which is divided into three consecutive segments each containing a variable number of samples, n, as illustrated in Figure 7.1.

$$\begin{aligned} x_n &= \{x(i) = s(i-n) \mid i \in 1, \ldots, n\} \\ y_n &= \{y(i) = s(i) \mid i \in 1, \ldots, n\} \\ z_n &= \{z(i) = s(i+n) \mid i \in 1, \ldots, n\} \end{aligned} \tag{7.1}$$

Frames at the beginning of an utterance, for which x_n is not fully defined, are classified as 'silent'; likewise frames at the end of an utterance, for which y_n and z_n are not fully defined. The use of the third segment of speech, z_n, is described later in this Section.

The analysis of the segments x_n and y_n produces an estimate of the *FØ* value for a frame of speech data centred $t_{interval}(r-1)$ milliseconds into the utterance. This timing

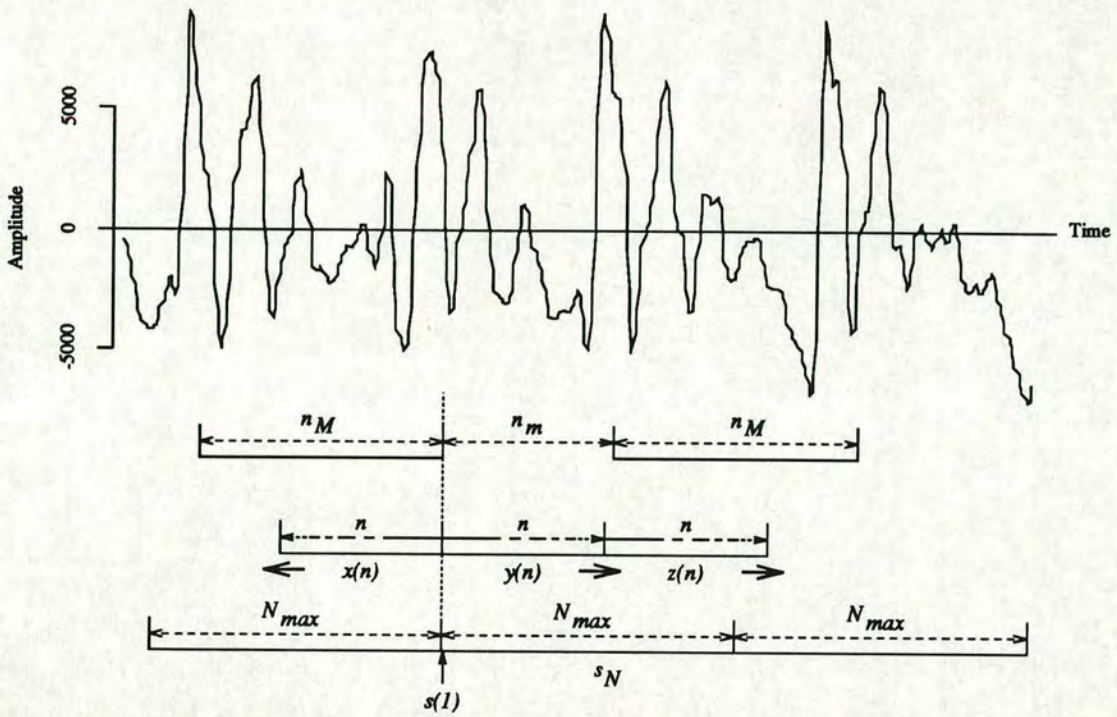


Figure 7.1: Analysis segments for the enhanced super resolution *FØ* determinator (*e*SRFD)

is independent of the estimated fundamental period of the analysis frame.

An *FØ* contour produced by any *FØ* determination algorithm can be expected to contain values which are inaccurate, such as instances of *FØ* doubling and halving errors and instances when voiced sections of speech are classified as unvoiced by an FDA or when unvoiced sections of speech are classified as voiced. *FØ* doubling errors occur when the estimated fundamental frequency is an overtone of the true fundamental frequency. *FØ* halving errors occur when the *FØ* determination algorithm erroneously mistakes the correctly estimated fundamental frequency as an overtone of the true fundamental frequency and then effectively over-compensates for this by dividing by some multiple of two. The remaining modifications made to the SRFD algorithm are designed to reduce the occurrences of *FØ* doubling and halving errors and to improve the voicing classification of frames of speech.

A value of n is sought for each analysis frame such that each segment defined in

Equation 7.1 occupies a single fundamental cycle. Candidate values of n are sought within the range N_{min} to N_{max} by using the normalised crosscorrelation coefficient $p_{x,y}$ defined in Equation 3.10. The locations of local maxima in $p_{x,y}$ with values above the adaptive threshold T_{srfd} (Equation 3.12) form candidate values for the fundamental period.

If no candidates for the fundamental period are found, the frame is classified as 'unvoiced'. Otherwise, the frame is classified as 'voiced' and a second normalised cross-correlation coefficient $p_{y,z}(n)$ is determined for all the fundamental period candidates.

$$p_{y,z}(n) = \frac{\sum_{j=1}^{\lfloor n/L \rfloor} y(jL).z(jL)}{\sqrt{\sum_{j=1}^{\lfloor n/L \rfloor} y(jL)^2 . \sum_{j=1}^{\lfloor n/L \rfloor} z(jL)^2}} \tag{7.2}$$

$$\{n \mid p_{x,y}(n) > T_{srfd}\}$$

Those candidates for which $p_{y,z}(n)$ also exceeds the threshold T_{srfd} are given a score of 2, while the others are given a score of only 1. Candidates with a higher score are more likely to represent the true fundamental period. If there are one or more candidates with a score of 2, then all those with a score of only 1 are removed from the list of candidates and ignored. Following this, if there is only one candidate (with a score of either 1 or 2) then the candidate is assumed to be the best (and only) estimate of the fundamental period for that frame. Otherwise, an optimal fundamental period is sought from the set of remaining candidates. The candidates are listed in order of increasing fundamental period. The candidate at the end of this list represents a fundamental period of n_M, and the m'th candidate represents a period of n_m. The coefficient $q(n_m)$ (Equation 7.3) is calculated for each candidate. $q(n_m)$ is the normalised crosscorrelation coefficient between two sections of length n_M spaced n_m apart, as illustrated in Figure 7.1.

$$q(n_m) = \frac{\sum_{j=1}^{n_M} s(j - n_M).s(j + n_m)}{\sqrt{\sum_{j=1}^{n_M} s(j - n_M)^2 . \sum_{j=1}^{n_M} s(j + n_m)^2}} \tag{7.3}$$

As in the unmodified version of SRFD, the first coefficient $q(n_1)$ is then assumed to be the optimal value. If a subsequent $q(n_m)$ exceeds this optimal value when multiplied by 0.77 (an *a priori* value) then it is in turn assumed to be the optimal value. The candidate for which $q(n_m)$ is believed to be the optimal value forms the estimate for the fundamental period, n_0, of the frame being analysed.

In the case when there is only one fundamental period candidate with a score of 1 and no candidates with a score of 2, there is only a small probability that the candidate correctly represents the true fundamental period of the frame. If, in such cases, the previous frame was classified as either 'silent' or 'unvoiced', then the *FØ* value describing the current 'voiced' frame is held until the state of the subsequent frame is known. If this next frame is also not classified as 'voiced', then the frame whose *FØ* value is on hold is an isolated frame which is highly unlikely to be voiced. It is therefore re-classified as 'unvoiced'. Otherwise, the held *FØ* value is assumed to be a sufficiently good estimate of the fundamental frequency for that frame.

The above modification dramatically reduces the occurrences of *FØ* doubling and halving errors in the resultant *FØ* contour. However, its implementation also causes a greater percentage of voiced regions of speech to be erroneously classified as unvoiced. In order to counteract this undesirable effect, an additional modification applies biasing to the coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ for values of n where the fundamental period of a new frame is expected to lie. Biasing is applied if the following conditions are satisfied:

- The two previously analysed frames were classified as 'voiced'.
- The *FØ* value of the previous frame is not being temporarily held.
- The fundamental frequency of the previous frame f_0' is less than $\frac{7}{4}$ times the fundamental frequency of its preceding voiced frame f_0'', and greater than $\frac{5}{8}f_0''$, ie. it is highly probable that the fundamental period estimate of the previous frame is not an *FØ* doubling or halving error.

The fundamental period of the new frame n_0 is expected to lie within the range of n closest to n_0' for which the set of $p_{x,y}(n)$ from the previous frame are greater than zero (Figure 7.2). The normalised crosscorrelation coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ are doubled for values of n in this range. This effectively applies a bias on the location of a maxima in

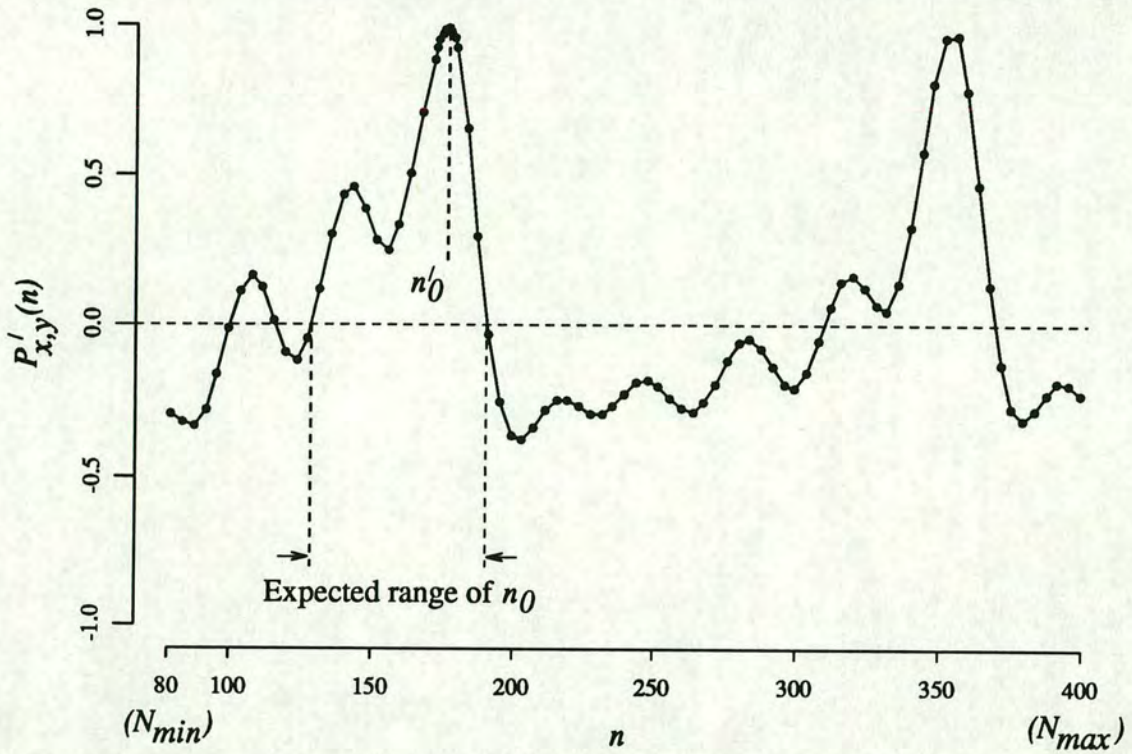


Figure 7.2: Example set of $p'_{x,y}(n)$ from previous frame

the region of the fundamental period for the previous frame to form a candidate for the fundamental period of the current frame. Note, however, that the decision to classify a frame of speech as voiced or unvoiced is based on the presence or absence of local maxima in $p_{x,y}(n)$ which exceed the adaptive threshold T_{srfd}. The biasing therefore tends to increase the percentage of unvoiced regions of speech being incorrectly classified as 'voiced'. In order to minimise this undesirable side effect, if the unbiased coefficient $p_{x,y}(n)$ does not exceed the threshold T_{srfd} for the candidate believed to be the best estimate of the frame fundamental period, then the $F\emptyset$ value for that frame is held until the state of the subsequent frame is known. If this next frame is classified as 'silent' or 'unvoiced', the former frame is re-classified as 'unvoiced'.

Finally, as in the unmodified version of SRFD, the algorithm obtains an estimate of the fundamental period with a fine resolution. A more accurate fundamental period estimate for the frame is determined by calculating $r_{x,y}(n)$ (Equation 3.9) for n in the

region $n_0 - L$ to $n_0 + L$. The location of the maximum within this range corresponds to a more accurate value of the fundamental period. This final estimate is then refined to eliminate the effect of time quantisation errors, by using an interpolation method which is described in Medan *et al.* (1991).

7.2 Evaluation of *FØ* determination algorithms

The following evaluation of the algorithms summarised in Section 3.2 and the enhanced SRFD algorithm described in the previous Section, is tailored to identify the most appropriate method of fundamental frequency extraction of adult male and female speech recorded with little ambient interference.

A database containing approximately five minutes of speech is used in the evaluation. The speech is recorded simultaneously with a close-talking microphone and laryngograph in an anechoic studio. The database is formed from fifty sentences each read by one adult male (♂) and one adult female (♀), both with non-pathological voices. The database is biased towards utterances containing voiced fricatives, nasals, liquids and glides, since these phones are aperiodic in comparison to vowels and FDAs generally find them difficult to analyse.

The sentences are recorded with the use of a laryngograph so that a reference laryngeal frequency (*Fx*) contour can be obtained. The laryngograph measures the impedance between two electrodes placed bilaterally across the larynx. The measured impedance decreases with an increasing degree of vocal-fold contact. Glottal closure is marked in the laryngograph signal (*Lx*) by a sharp rise to a peak, whereas the opening of the glottis is much slower and less marked by a gradual fall in signal amplitude (Baer *et al.*, 1983). It is not expected that the laryngograph data can be used to give a perfectly 'correct' corresponding *Fx* contour for the associated speech data. In particular, a contour generated from laryngograph data may be inaccurate due to effects at the end of voiced speech segments for which a small area of vocal-fold contact is insufficient for the glottis activity to be distinguished from noise in the laryngograph signal, but the speech is periodic and low in energy. The extent of such errors will only be over two or three *Fx* cycles and are thus deemed negligible in this study. The laryngograph data provides a simple and (relative to FDAs or locating glottal closures from the speech waveform by hand)

accurate method of producing a contour with which other contours can be compared.

The functionality of all the algorithms in the evaluation is dependent upon certain thresholds and pre-determined parameters, some of which are common across the algorithms. In order to set a degree of similarity between the FDAs, all are required to present a computed *FØ* value at 6.4ms intervals through an utterance. The values are limited to the ranges of 50Hz–250Hz for the male speaker and 120Hz–400Hz for the female speaker. In cases where a fixed-length analysis frame is required by an algorithm, the frame duration has been set to 38.4ms. A frame of this duration enables at least two signal periods to reside within the frame for all fundamental frequencies greater than 52Hz, and allows sufficient data for cepstral and spectral analysis techniques. The speech data applied to the algorithms is sampled at 20kHz using a 16-bit analogue-to-digital converter (ADC). For those algorithms which operate on a low-pass filtered waveform, the low-pass filtered speech is produced by a finite impulse response (FIR) filter with a -3dB cut-off at 600Hz and rejection greater than -85dB above 700Hz, though these filter characteristics are not crucial to the performance of the FDAs.

7.2.1 A laryngeal frequency tracker

The reference *Fx* contours are created from the laryngograph data by using a simple pulse location algorithm and deriving the duration between successive pulses. Assume that the laryngograph data contains a positive-going pulse at the instant of each glottal closure and that any acoustic propagation delay between the glottis and the close-talking microphone (approximately 0.5ms) has been compensated.

A 'pulse' is defined as follows (see Figure 7.3): The pulse start time t_{start} is the first sample for which the amplitude is less than zero and less than or equal to the amplitude of following samples. The pulse stop time t_{stop} is the last sample for which the amplitude is less than zero and less than or equal to the amplitude of preceding samples. The pulse width t_{width} is defined as the difference between t_{start} and t_{stop}. The pulse peak amplitude a_{peak} is the maximum amplitude of samples between t_{start} and t_{stop} (always greater than zero). The pulse instant t_{pulse} is defined as the time of the first of these samples with an amplitude a_{peak}.

For laryngograph data sampled at 20kHz by a 16-bit ADC, a pulse at t_{pulse} is classed

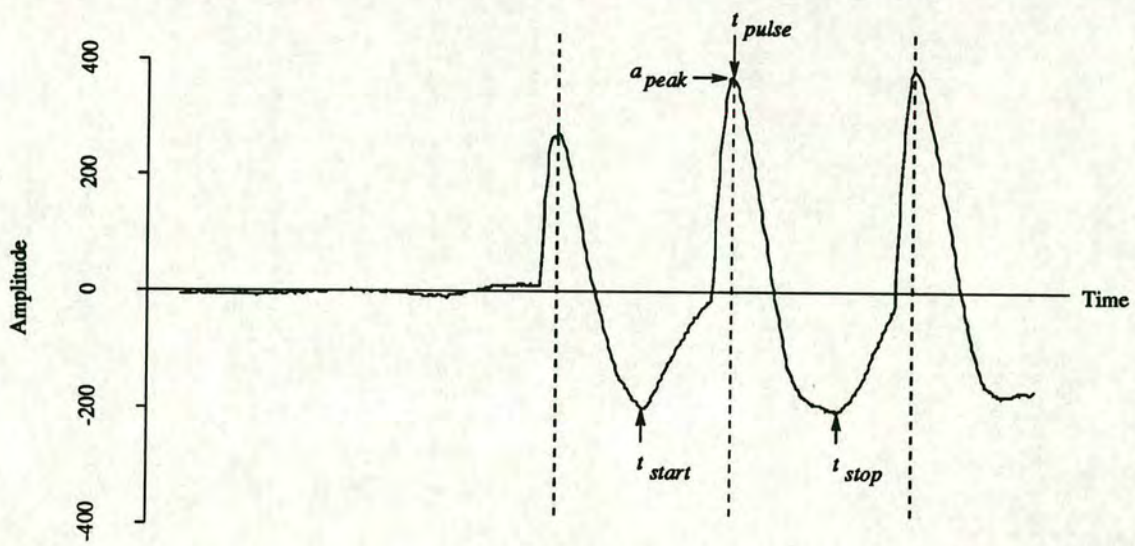


Figure 7.3: Glottal pulses evident in laryngograph data

as a marker of the glottal closure instant if the pulse width t_{width} is greater than four samples (for data from both speakers) and the pulse peak amplitude a_{peak} is greater than a threshold value which is dependent on the recording conditions and which relates to the signal-to-noise level in Lx.

The set of pulse instances generated from the laryngograph data of an utterance are considered in chronological order. The duration between one pulse instant t^n_{pulse} and the next pulse instant t^{n+1}_{pulse} is calculated and converted to Hertz. If the value is greater than some lower limit then it is taken to represent the laryngeal frequency at the time $(t^n_{pulse} + t^{n+1}_{pulse})/2$; otherwise, the duration between the pulses is considered to correspond to an unvoiced region of speech. The Fx values are limited to $\geq$50Hz for the male speaker, and $\geq$120Hz for the female speaker. There must be at least three laryngograph pulses in each voiced section. This final restriction is imposed in order to remove the few errors when a 'pulse' in the laryngograph data is formed by events other than glottal activity.

The Fx determined using this algorithm has a mean of 124.19Hz and a population standard deviation of 25.61Hz for the male speech, and a mean of 256.02Hz and standard

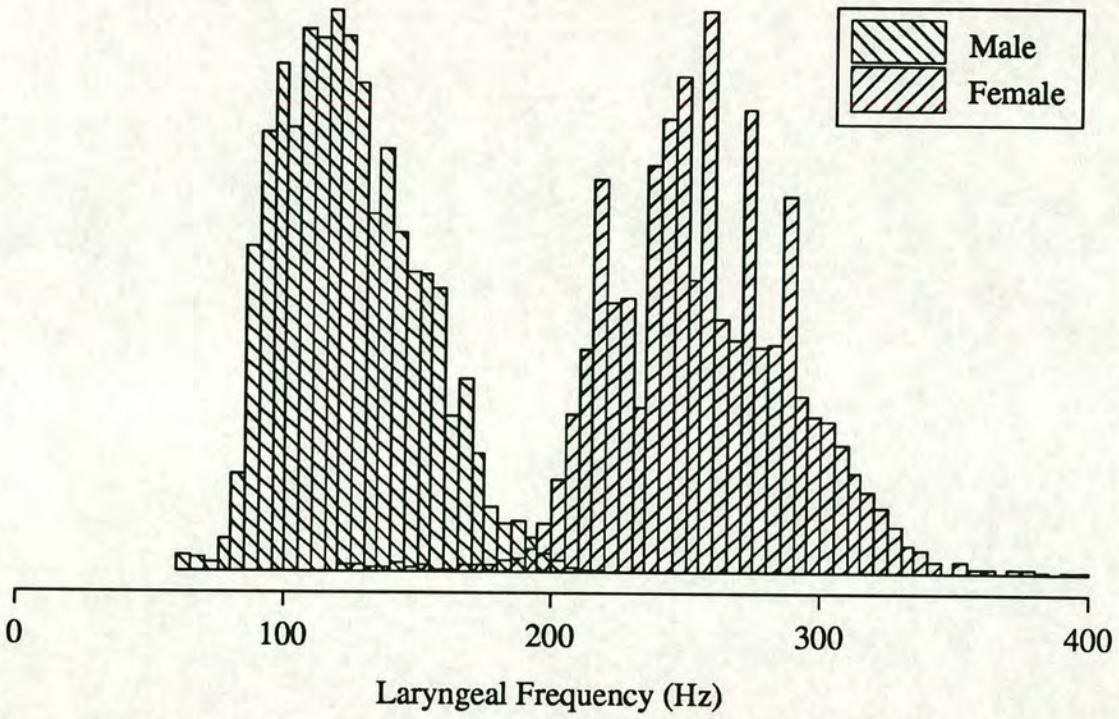


Figure 7.4: Histograms of laryngeal frequency for male and female speech

deviation of 33.39Hz for the female speech. These statistics indicate that the selection of ranges for which $F\emptyset$ is to be determined by the FDAs (♂: 50–250Hz; ♀: 120–400Hz) is suitable. The histograms of laryngeal frequency in Figure 7.4 for the male and female speech both show a single prominent lobe of data centred around the mean. There are no smaller lobes to the left or right of the main one, suggesting that Fx doubling and halving is not a problem for the algorithm.

The accuracy with which Fx can be determined by this method is limited by the time quantisation in sampling the laryngograph signal. The pulse instances can only be specified to within a single sample duration. The true laryngeal period T_x estimated by the duration between adjacent pulses d samples apart lies within the range,

$$(d-1).T_s < T_x \leq d.T_s \tag{7.4}$$

where T_s is the sampling period. The true laryngeal period can be specified with an

error of just T_s seconds — note that the error is independent of d. The true laryngeal frequency $F_x = 1/T_x$ lies within the range,

$$\frac{F_s}{(d-1)} < F_x \leq \frac{F_s}{d} \tag{7.5}$$

where F_s is the sampling frequency. Each value of F_x therefore has an error of,

$$\frac{F_s}{d(d-1)} \text{ Hz} \tag{7.6}$$

This error is dependent upon d and is calculated for the distributions of Fx for the two speakers. The quantisation error in determining Fx by this method has a mean of 0.80Hz and population standard deviation of 0.34Hz for the male speaker, and a mean of 3.33Hz and standard deviation of 0.86Hz for the female speaker. This error cannot be compensated for and affects the evaluation results for the FDAs.

7.2.2 Comparison of asynchronous frequency contours

The reference Fx contour of each sentence, generated from the laryngograph data, is compared with the $F\emptyset$ contours generated by the six algorithms reviewed in Section 3.2 and the enhanced $F\emptyset$ determination algorithm eSRFD described in Section 7.1, all of which operate on the speech data.

Frequency contours are described by a sequence of two variables — a time from the start of the utterance and a frequency value at that time. The frequency value is set to zero to describe regions of silence or unvoiced speech. For the reference Fx contour, the times from the start of an utterance (where an Fx value is stated) are given by the mid-point times between pulses detected in the laryngograph data. The laryngeal frequency between such times must be inferred by interpolation. Similarly, for the $F\emptyset$ contours generated by the FDAs in which an $F\emptyset$ value is only stated at regular intervals; linear interpolation is used to form an approximation of the $F\emptyset$ value between times for which it is stated.

For each time a frequency value is stated in either the reference Fx contour or the $F\emptyset$ contour, the two values for the contours (one of which may be derived by linear interpolation) are compared in the following manner (see Figure 7.5). The frequency

value from the reference contour is represented by $Fx_{reference}$, and that from the $F\emptyset$ contour is represented by $F0_{FDA}$.

- When $Fx_{reference}$ and $F0_{FDA}$ are zero, both contours describe a silent or unvoiced region of the utterance and no error results.
- If $F0_{FDA}$ is non-zero but $Fx_{reference}$ is zero, then an unvoiced (or silent) region of speech has been incorrectly classified as voiced by the FDA. The duration of the erroneous region is determined by finding the subsequent time at which either the $F\emptyset$ contour becomes zero or the reference Fx contour becomes non-zero.
- If $Fx_{reference}$ is non-zero but $F0_{FDA}$ is zero, then a voiced region of speech has been incorrectly classified as unvoiced (or silent) by the FDA. The duration of the region in error is determined by finding the subsequent time at which either the reference Fx contour becomes zero or the $F\emptyset$ contour becomes non-zero.
- When both $F0_{FDA}$ and $Fx_{reference}$ are non-zero, the contours represent the correct classification of voiced speech. In such cases, if the ratio of,

$$\frac{Fx_{reference} - F0_{FDA}}{Fx_{reference}} \tag{7.7}$$

is greater than 0.2 then the FDA made a gross error in estimating the fundamental frequency of more than 20% of the reference Fx and the error is categorised as $F\emptyset$ 'halving'. If the ratio in Equation 7.7 is less than -0.2, a gross $F\emptyset$ error of more than 20% was made which is categorised as $F\emptyset$ 'doubling'[1]. Otherwise, the FDA is assumed to have estimated the fundamental frequency with an acceptable accuracy, and the absolute difference between $F0_{FDA}$ and $Fx_{reference}$ is noted. The $\pm$20% threshold of acceptability is chosen because all FDAs are expected to form an $F\emptyset$ value within this range with due consideration of time quantisation errors and the finite frequency resolution of the analysis technique.

The durations of unvoiced or silent regions classified in error and the durations of voiced sections incorrectly classified as unvoiced or silent by the FDA, are accumulated

[1] $F\emptyset$ 'halving' and $F\emptyset$ 'doubling' terms are used here to distinguish between abnormally low and abnormally high estimates of $F\emptyset$ rather than referring to exact octave errors.

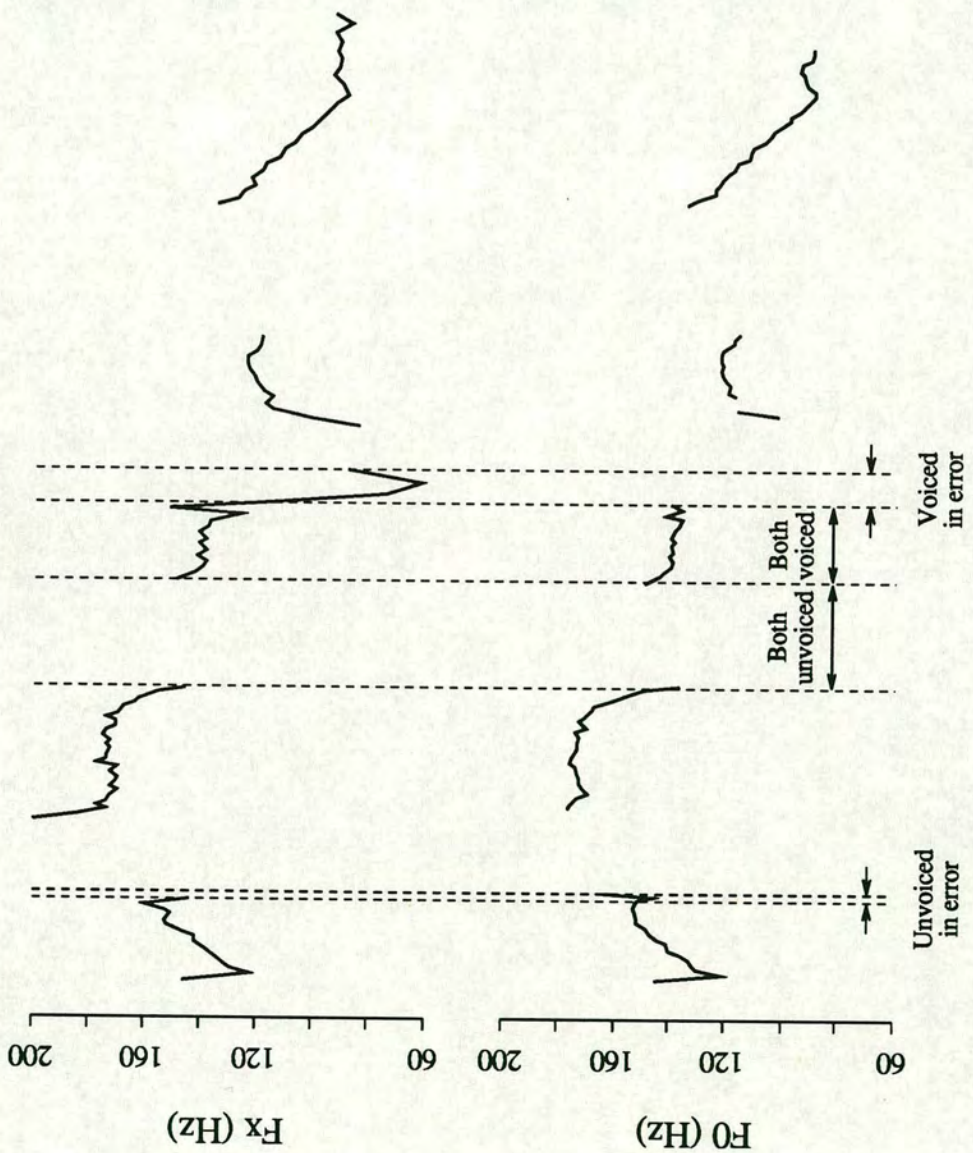


Figure 7.5: Comparison of asynchronous Fx and $F\emptyset$ contours

over all the utterances in the database for each speaker. The sum of the individual durations of such erroneous regions are expressed as a percentage of the total duration of unvoiced (or silent) speech and voiced speech respectively. The total number of comparisons for which a gross *FØ* doubling error occurs and the total number of gross *FØ* halving errors are also determined. These gross *FØ* error rates are expressed as a percentage of the total number of comparisons for which $F0_{FDA}$ and $Fx_{reference}$ are non-zero. The final statistics calculated in the evaluation of an FDA are the population standard deviation (p.s.d.) and the average, absolute deviation of the reference *Fx* and *FØ* contours when both represent voiced speech, and the FDA has not made a gross *FØ* error.

7.2.3 Results and discussion

Each of the six FDAs reviewed in Section 3.2 and the enhanced SRFD algorithm described in Section 7.1 is used to extract the fundamental frequency from the speech waveforms of the utterances in the database. The contours generated are then compared with the reference *Fx* contours formed from synchronised laryngograph data, using the method described above. The results of the comparison of each pair of contours are averaged over the entire set of utterances for each speaker so that the evaluation of the FDAs may be assumed to be independent of the segmental and prosodic structures of the read utterances. The resultant statistics are presented in Figures 7.6 & 7.7 for the two speakers.

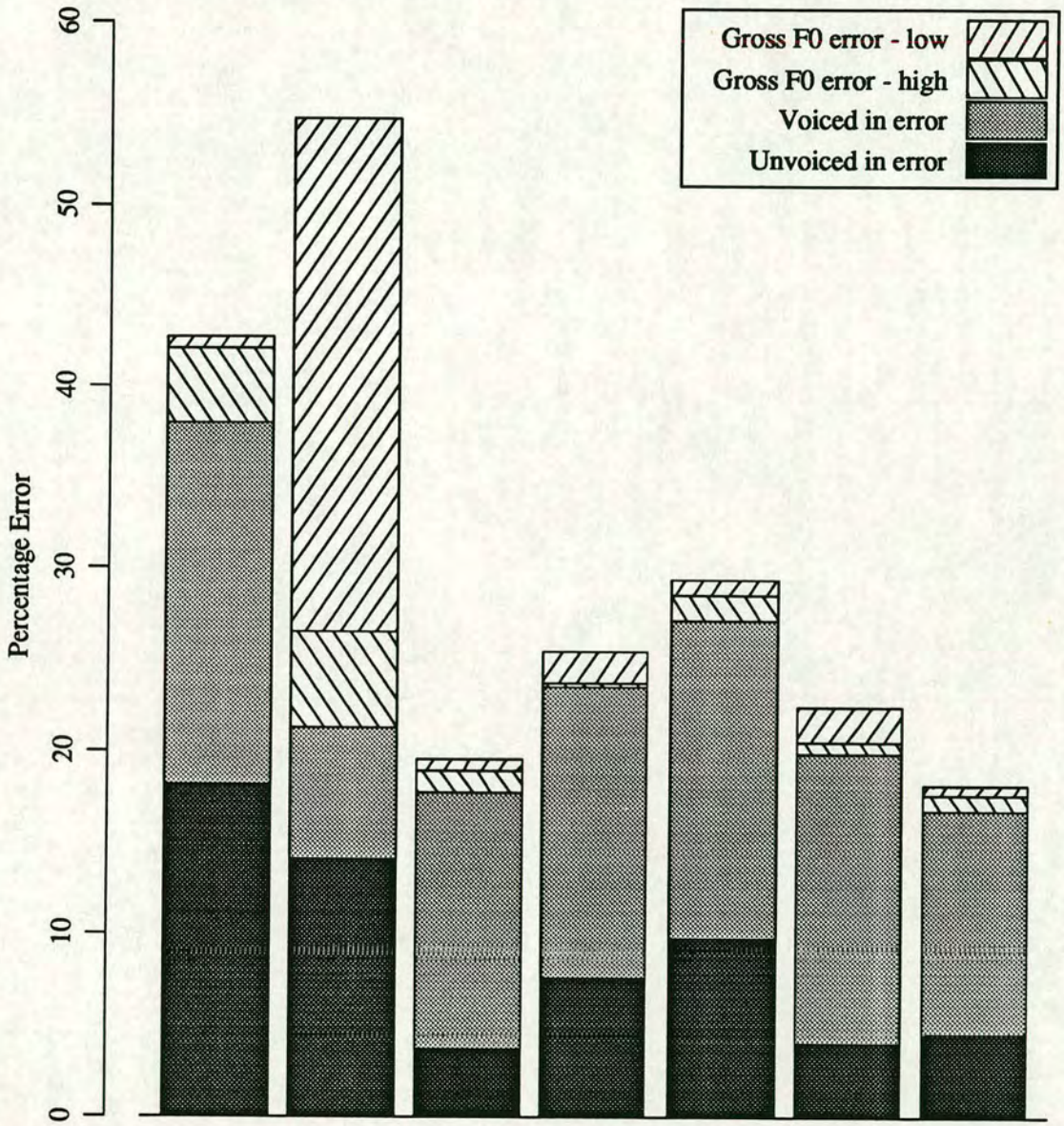
Voiced/Unvoiced Classification: CFD and HPS (which use a frequency domain representation of speech) perform considerably poorer for the female speech than for the male speech when classifying frames of speech as either voiced or unvoiced. Results reported by Rabiner *et al.* (1976) also find the cepstral-based *FØ* determination to perform much better on lower frequency speech than on higher frequency speech. This is a consequence of the frequency domain algorithms basing the voicing classification on the magnitude of spectral peaks rather than using the temporal structure of the waveform. The maximum cepstral peak magnitude in a given frame is a measure of the degree of voice periodicity; however, voiced speech is only approximately periodic. During sudden changes in articulation, the speech waveform can be aperiodic although voiced, for example in the

transition from a vowel to a nasal segment. In such cases, a single feature such as the cepstral peak magnitude is insufficient to distinguish voiced speech from unvoiced speech (Atal & Rabiner, 1976). The time domain approaches, FBFT, PP, IFTA and SRFD, are consistently better, with both FBFT and *e*SRFD having an overall voicing decision error rate of less than 17%.

Gross F*Ø Errors:* The outstanding number of *FØ* halving errors produced by HPS for the male speech is unacceptably high. Large sections of aperiodic voiced speech are often analysed by the HPS algorithm to have an erroneously low fundamental frequency. For the male speech, CFD and HPS generate far more *FØ* doubling errors than the time domain based algorithms. The total gross *FØ* error rates for the time domain based algorithms are less than 2.1% for the male speech and less than 4.2% for the female speech, with the exception of the unmodified SRFD which generates 5.6% *FØ* halving errors for the female speech. This error rate is dramatically reduced to 0.2% by the modified algorithm *e*SRFD.

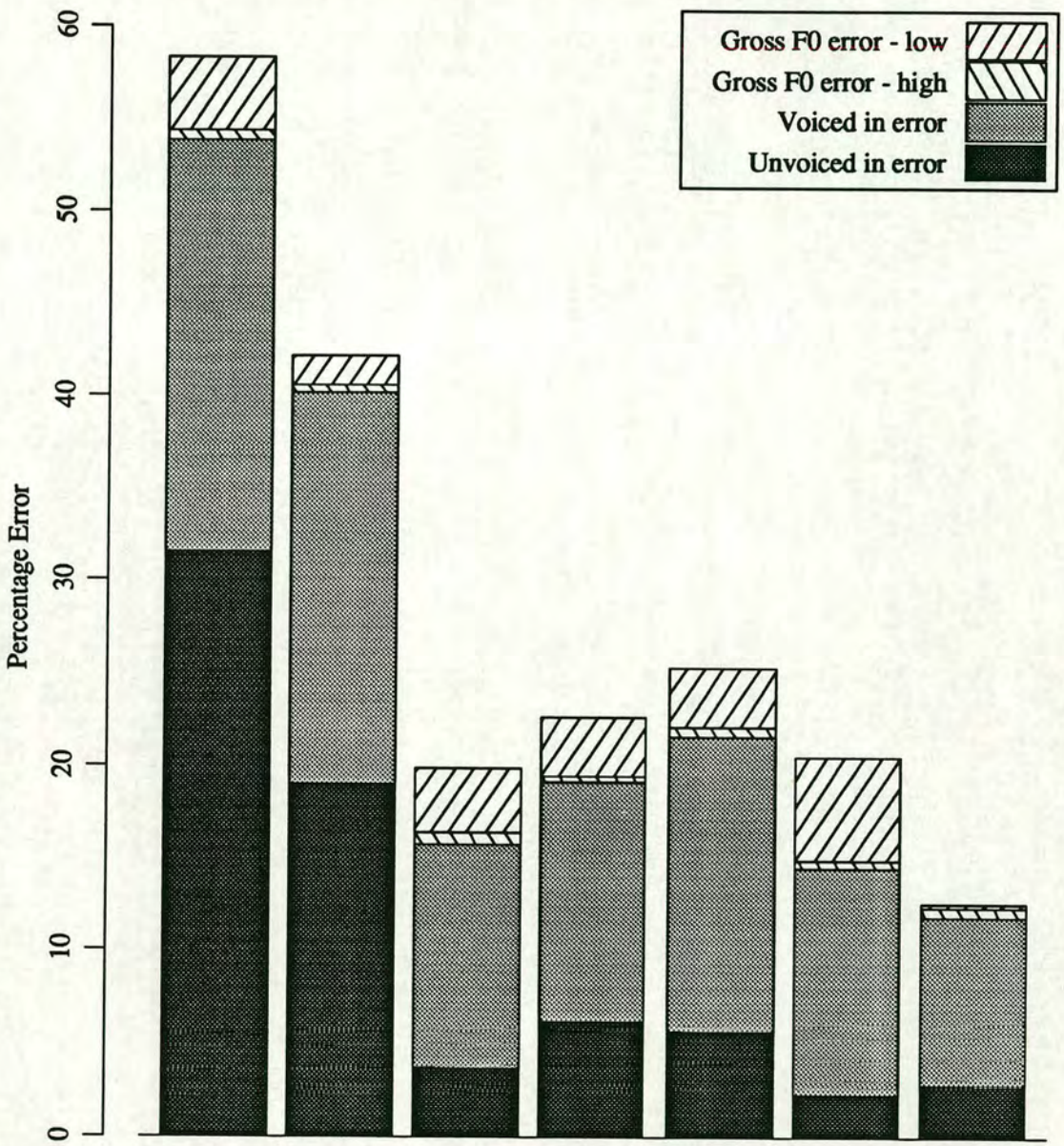
Frequency Accuracy: The accuracy with which the fundamental frequency is determined by the FDAs is not of great importance for the analysis of intonation since only the general trend of *FØ* is required. However, it can be seen that the super resolution *FØ* determinator performs consistently better that the other FDAs in this category, with the accuracy of its contours being comparable with that of the reference *Fx* contours.

Any FDA producing an *FØ* contour suitable for automatic prosodic analysis must perform consistently between male and female speech. The resultant contour must accurately identify voiced sections of speech so that pitch accents are not left undetected, and gross *FØ* errors must be minimal to facilitate correction by *FØ* post-processing techniques. CFD and HPS are therefore unsuitable algorithms for the application of automatic prosodic analysis. Of the remaining FDAs studied here, FBFT and *e*SRFD form the better voiced/unvoiced classifications and generate the least number of gross *FØ* errors. The enhanced super resolution *FØ* determination algorithm (*e*SRFD) is used to generate the fundamental frequency contours for the automatic prosodic analysis described in this thesis.



|Fx-F0|	CFD	HPS	FBFT	PP	IFTA	SRFD	eSRFD
mean (Hz)	2.94	3.25	1.86	2.64	2.67	1.78	1.40
p.s.d. (Hz)	3.60	3.21	2.89	3.01	3.37	2.46	1.74

Figure 7.6: FDA evaluation: Male speech



|Fx-F0|	CFD	HPS	FBFT	PP	IFTA	SRFD	eSRFD
mean (Hz)	6.39	4.59	5.40	6.11	4.38	4.14	4.17
p.s.d. (Hz)	7.61	5.31	7.03	6.45	5.35	5.51	5.13

Figure 7.7: FDA evaluation: Female speech

7.3 *FØ* contour post-processing

An attempt is made in the design of the *FØ* extraction algorithm (*e*SRFD) to minimise the number of *FØ* doubling and halving errors derived from the speech waveform. It is, however, inevitable that some errors of this kind will exist in the *FØ* contour and that they will propagate into subsequent analysis of the contour. The aim of post-processing an *FØ* contour is to eliminate *FØ* doubling and halving errors whilst maintaining any information in the *FØ* contour which is relevant to the prosodic analysis. Whereas the previous attempt (Section 7.1) to reduce the number of errors was based on the characteristics of the speech waveform, the post-processing of the contour aims to reduce the number of errors by taking into account the characteristics of the contour itself.

7.3.1 Non-linear smoother

In an attempt to eliminate *FØ* doubling and halving errors, the contour can be processed by a non-linear filter. Linear smoothing can be applied to remove small perturbations in the contour which are not required in the prosodic analysis. Linear smoothing also reduces the effects of frequency quantisation in the contour. The non-linear filter and linear smoothing proposed by Rabiner *et al.* (1975) is illustrated in Figure 7.8.

A non-linear median filter consists of a shift register of length l_{median}. The *FØ* values (one for each analysis frame) stored in the elements of the register are sorted into ascending order and the middle value is presented at the output of the median filter. The length l_{median} is always an odd number.

A linear Hann window is used as a smoothing filter. The window uses l_h values selected from a shift register of length l_{Hann} and weights the n'th value by a factor $h(n)$.

$$h(n) = \frac{1}{l_h + 1}\left(1 - \cos\frac{2\pi n}{l_h + 1}\right), \qquad n \in 1, \ldots, l_h \tag{7.8}$$

If the shift register contains more than $l_{Hann}/2$ *FØ* values which represent frames of unvoiced speech or silence, then the output of the filter is also set to represent unvoiced speech. Breaks in an *FØ* contour are represented by values equal to zero, which correspond to frames of unvoiced speech and frames of silence. Otherwise, l_h is set to the number of *FØ* values in the shift register which represent frames of voiced speech and

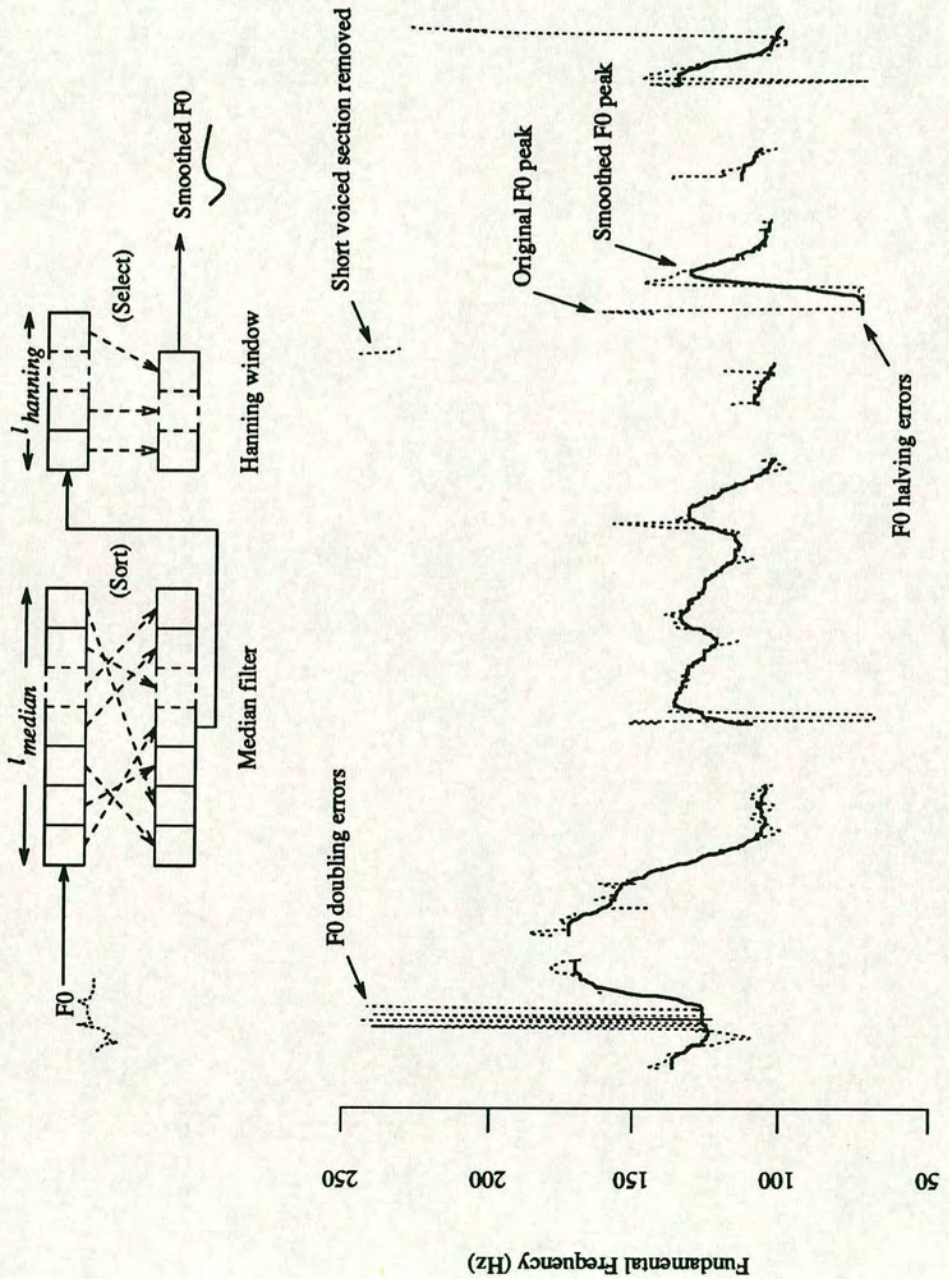


Figure 7.8: Non-linear smoother

the filter output is set to the sum of their weighted values.

The algorithm uses extrapolation for the initial and final values of the contour in order to prevent small time shifts due to the delay imposed by the smoothing process. By using a Hann window for which l_{Hann} is an odd number, the delay $(l_{median} + l_{Hann} - 2)$ imposed by the smoother is always an even number. The extrapolation at the beginning of the track accounts for half of this delay, and extrapolation at the end accounts for the other half.

An example of the effects of this non-linear smoothing process is illustrated in Figure 7.8. In this example, the number of elements in the shift register of the median filter and the maximum number of elements in the Hann window are set to 17 frames. Although some of the short-term irregularities of the *FØ* contour which are not relevant to prosodic analysis are removed by this process, some of the inflections of the *FØ* contour which need to be preserved for prosodic analysis are affected. There are two major disadvantages in applying a non-linear smoother to an *FØ* contour prior to prosodic analysis. Firstly, short isolated regions of the *FØ* contour which represent sections of voiced speech spanning less than $(l_{median} + 1)/2$ frames are removed if they are flanked by sequences of unvoiced speech each spanning $(l_{median} + 1)/2$ frames or more. Secondly, peaks in the *FØ* contour related to pitch accents are time-shifted and altered in magnitude if they occur near *FØ* doubling or halving errors. If the numbers of elements in the shift registers are reduced in order to preserve *FØ* inflections relevant to the prosodic analysis then fewer *FØ* doubling and halving errors are removed. If the numbers of elements in the shift registers are increased in order to remove larger sections of an *FØ* contour which are subject to doubling and halving errors then the magnitudes of *FØ* peaks are reduced. Changing the magnitudes of the *FØ* peaks is undesirable because they are needed for the prosodic analysis. Another method of removing *FØ* doubling and halving errors is required so that the number of elements in the shift registers can be reduced, thus enabling relevant *FØ* inflections to be preserved whilst still being able to smooth *FØ* micro-perturbations which are irrelevant to the prosodic analysis.

7.3.2 De-step *FØ* filter

An algorithm is proposed in this Section which eliminates *FØ* doubling and halving errors in order to overcome the disadvantages inherent to the non-linear smoother. The algorithm is illustrated in Figure 7.9.

It is necessary to distinguish legitimate frame-to-frame variations of the *FØ* contour from changes which are due to errors in estimating the fundamental frequency. Once this distinction is made, the *FØ* changes which correspond to doubling and halving errors can be corrected. It is assumed that *FØ* can legitimately increase or decrease from frame-to-frame by up to 75% during continuously voiced sections of speech, and that *FØ* can change by any amount across unvoiced sections of speech. An *FØ* increase from one frame to the next of more than 75% is defined for this algorithm to be an *FØ doubling* (exact doubling is a 100% increase) superimposed on a legitimate decrease of up to 25%. Similarly, an *FØ* decrease from one frame to the next of more than 75% is defined to be an *FØ halving* superimposed on a legitimate increase of up to 25%.

Consider each section of voiced speech which is uninterrupted by any frames of unvoiced speech. The onset *FØ* value of a section of voiced speech can correspond to an *FØ* doubling or halving error, or can be a legitimate value. A technique is required to determine which of these states the *FØ* value is in. Each *FØ* value in the section of voiced speech is therefore placed into one of a number of pools. The first *FØ* value is placed into pool P_x. If the change in *FØ* from one frame to the next corresponds to an *FØ* doubling (in terms of the definition above) then all *FØ* values up to the next *FØ* doubling or halving are placed into a higher pool (x is increased). Similarly, if the change in *FØ* corresponds to an *FØ* halving then all *FØ* values up to the next *FØ* doubling or halving are placed into a lower pool (x is decreased). The pool containing the greatest number of values is assumed to contain legitimate *FØ* estimates. This assumption is valid only if an FDA makes a majority of the *FØ* estimates without doubling or halving errors in any given section of voiced speech. Let $x = 0$ for the pool containing the greatest number of values. Higher pools ($x > 0$) contain *FØ* values which are doubling errors and lower pools ($x < 0$) contain *FØ* values which are halving errors.

To describe this explicitly in mathematical terms, let f_i represent the *FØ* value of the i'th frame in a section of voiced speech. The first *FØ* value f_1 is placed into pool $P_{x(1)}$.

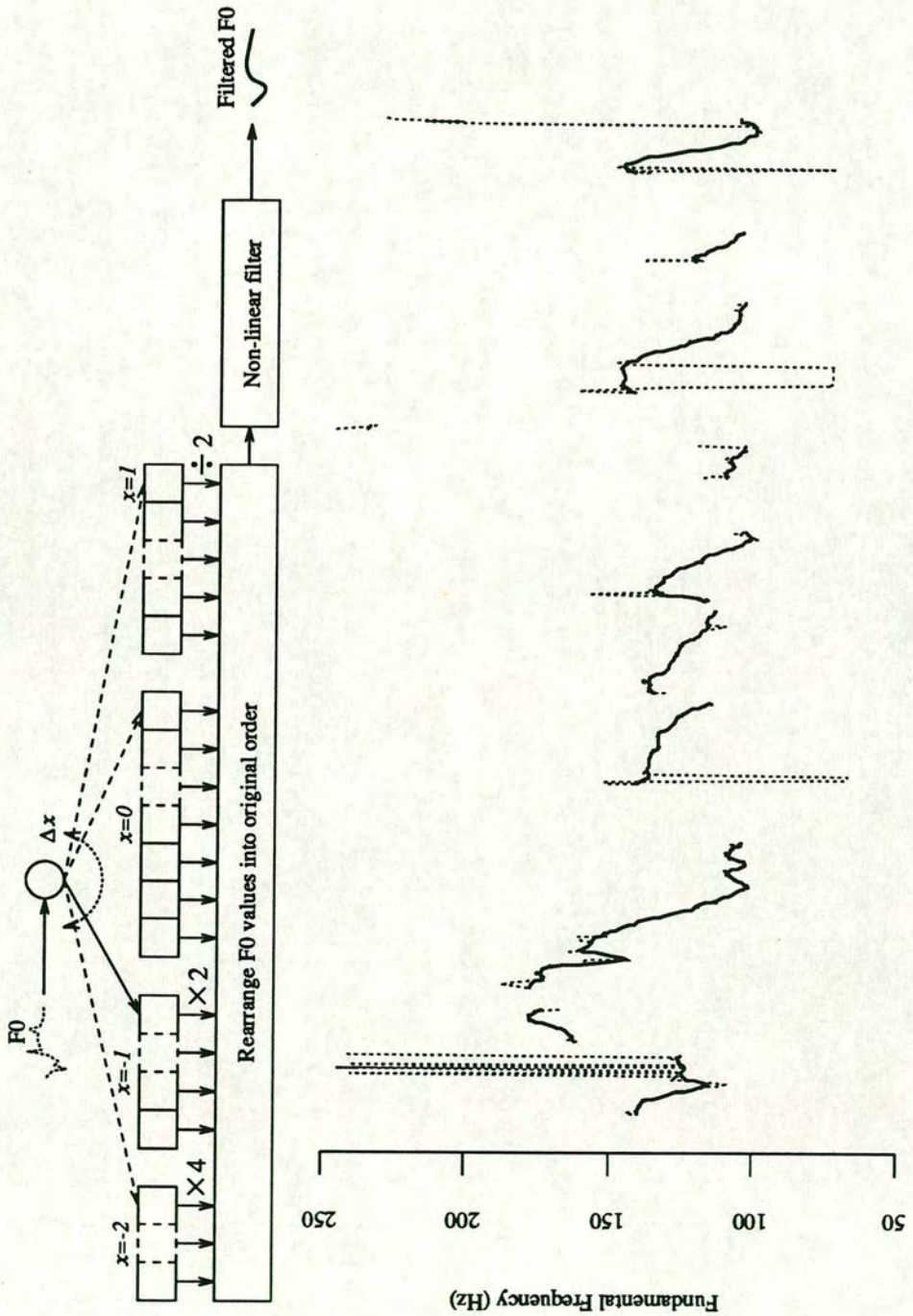


Figure 7.9: A de-step *FØ* filter and non-linear filter

Subsequent values ($i > 1$) are placed into a pool $P_{x(i)}$ where $x(i)$ is given in Equation 7.9. $F\emptyset$ values are pooled with the index i in order to keep track of their location in the $F\emptyset$ contour.

$$x(i) = \begin{cases} x(i-1) + \left\lfloor \log_2 \left(\frac{4}{7} . \frac{f_i}{f_{i-1}} \right) + 1 \right\rfloor & \text{if } f_i \geq f_{i-1} \\ x(i-1) - \left\lfloor \log_2 \left(\frac{4}{7} . \frac{f_{i-1}}{f_i} \right) + 1 \right\rfloor & \text{if } f_i < f_{i-1} \end{cases} \tag{7.9}$$

Let P_0 represent the pool containing the greatest number of $F\emptyset$ values. The $F\emptyset$ values in $P_{x(i)}$ are corrected by multiplying each value by $2^{-x(i)}$.

$$f_i' = 2^{-x(i)} f_i \qquad \forall f_i \in P_{x(i)} \tag{7.10}$$

Once the correction factor has been applied, the $F\emptyset$ contour for the voiced section is reconstructed from the new $F\emptyset$ values f_i' and the stored indices i.

An example of the effects of this de-step $F\emptyset$ filter on the performance of the non-linear smoothing process is illustrated in Figure 7.9. In this example, the number of elements in the shift register of the median filter and the maximum number of elements in the Hann window are set to 5 frames. The $F\emptyset$ doubling and halving errors are eliminated without affecting inflections of the $F\emptyset$ contour relevant to the prosodic analysis. $F\emptyset$ micro-perturbations are reduced by applying the non-linear smoother with a reduced number of shift register elements (reduced from spanning 17 frames to 5 frames). Short sections of voiced speech are therefore preserved for further analysis.

Some errors in the $F\emptyset$ contour can remain after this post-processing technique has been applied. If, in the course of a single voiced section of speech, an $F\emptyset$ doubling error is detected but a subsequent $F\emptyset$ halving error is not detected (or visa versa) then legitimate $F\emptyset$ values are altered. This produces doubling and halving errors rather than eliminating them. Legitimate $F\emptyset$ values are corrupted in this way only when a doubling error occurs during a legitimate rapid decrease in $F\emptyset$ such that the overall frame-to-frame increase is less than 75%, or when a halving error occurs during a legitimate rapid increase in $F\emptyset$ such that the overall frame-to-frame decrease is less than 75%. The possibility of this undesirable effect occurring is small and is far outweighed by the benefits of the de-step $F\emptyset$ filter.

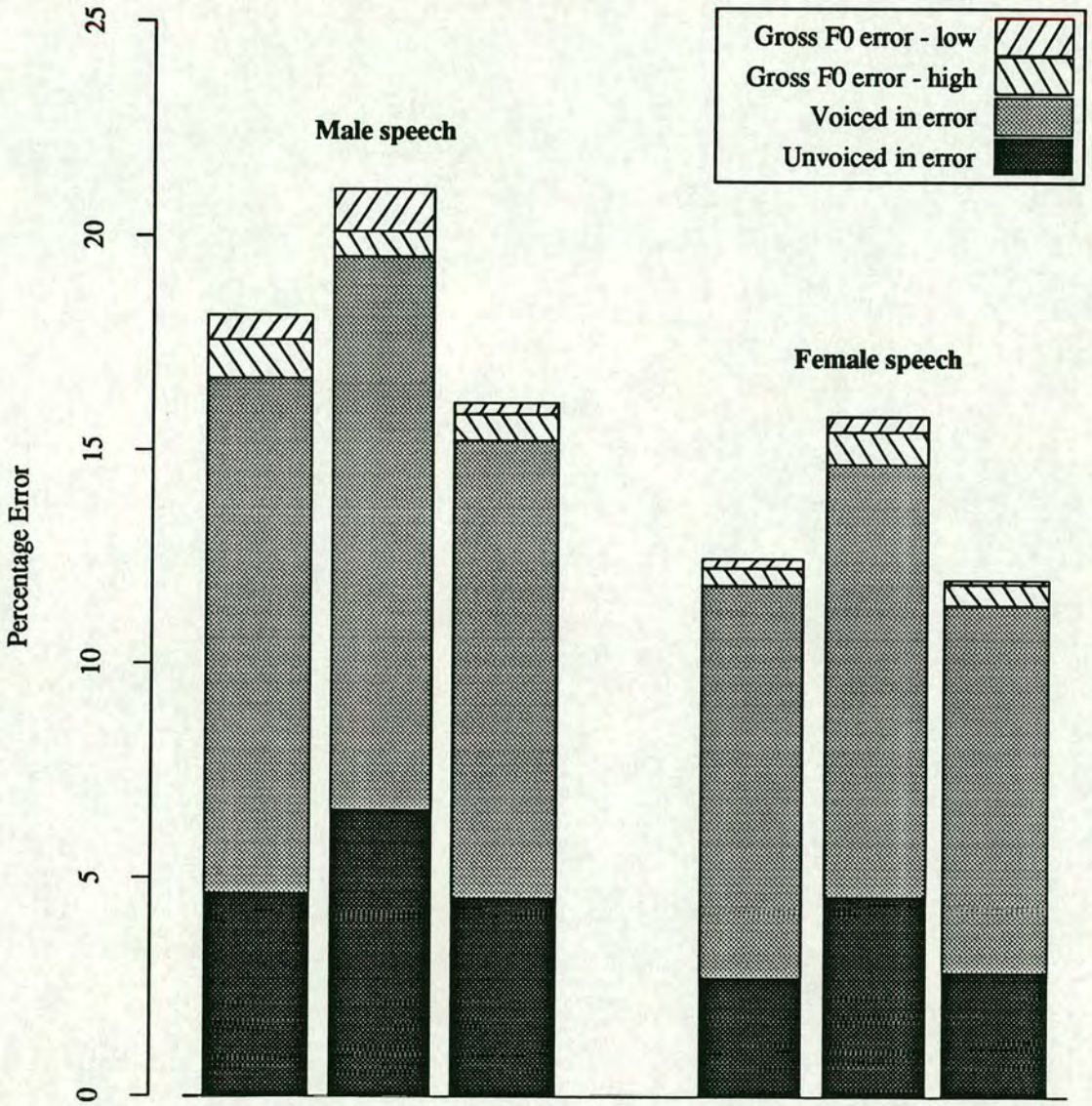
Three other types of error can remain in the $F\emptyset$ contour. Firstly, if an FDA makes

either more *FØ* doubling errors or more *FØ* halving errors than correct estimates of the fundamental frequency (in any given section of voiced speech) then pool P_0 will not contain legitimate values, as assumed. The *FØ* values in the voiced section will therefore be erroneously set to an overtone of the their true values. The only way to remove such errors is to improve the *FØ* determination algorithm. Secondly, the *FØ* contour may contain rapid frame-to-frame changes in frequency of more than 75% which are accurate estimates of the rate of glottal pulses, but which are detected as doubling or halving errors — such as in creaky voice. Finally, *FØ* inflections dependent on the segmental content of the utterance remain in the contour (such as *FØ* peaks at the onset of vowels occurring after unvoiced stop consonants) although they are not required by the prosodic analysis. A further level of post-processing is required, such as *FØ* contour stylisation (Sections 3.4 & 8.2.1), to eliminate these microprosodic variations in the *FØ* contour.

An evaluation of the *FØ* contours produced by the *e*SRFD algorithm when post-processed by the non-linear smoother and the de-step filter are shown in Figure 7.10. This shows that the non-linear filter only decreases *FØ* doubling errors in the male speech, while the de-step filter combined with a non-linear filter using fewer frames yields improvements in voicing classification and a reduction in gross *FØ* errors. These improvements are made at the expense of a slight reduction in the accuracy of *FØ* estimates.

7.4 Summary

Intonation refers to the manipulation of pitch for linguistic and paralinguistic purposes above the level of phonemic segments. The acoustic correlate of pitch is the fundamental frequency of speech. The extraction of *FØ* from a speech signal is therefore a prerequisite of prosodic analysis. A number of modifications to the SRFD algorithm (summarised in Section 3.2.6) are proposed in this Chapter, which reduce the occurrence of errors involved in the extraction of *FØ* such that it is optimised for prosodic analysis, relative to a selection of other algorithms designed to determine *FØ*. A novel de-step filter is proposed to post-process an *FØ* contour generated by an FDA in a way which further reduces the occurrence of discontinuity errors commonly observed in *FØ* contours.



|Fx-F0|	eSRFD	Post Smoother	Post De-step	eSRFD	Post Smoother	Post De-step
mean (Hz)	1.40	3.50	1.60	4.17	6.25	4.22
p.s.d. (Hz)	1.74	4.09	2.22	5.13	7.29	5.33

Figure 7.10: Evaluation of post-processed *FØ* contours (for *e*SRFD)

Chapter 8

Automatic Prosodic Analysis

The features of duration and energy, and the fundamental frequency extracted from the speech waveform need to be abstracted to form an acoustic-phonetic representation of sentential stress patterns and intonation in a syllabic domain. This is the aim of the automatic prosodic analysis system proposed here.

The syllabification of speech forms a central part of automatic prosodic analysis, taking on three roles. Firstly, the prosodic aspects of speech are described in a syllabic domain. Stress refers to the relative perceptual prominence of syllables and the description of intonation involves the association of pitch accents with prominent syllables. Secondly, the extraction of acoustic parameters is dependent upon the definition and identification of syllables. The duration feature found to optimally correlate with stress in Chapter 4 is dependent upon identifying the syllable lhymes in an utterance. The optimal energy feature determined in Chapter 5 is dependent upon identifying the syllable nuclei in an utterance. The third role of the syllabic domain is related to the integration of the acoustic parameters. The syllabic domain inherently encompasses energy and segmental information which can be passed on to the analyses of energy, duration and *FØ* measures. For example, the location of syllable nuclei can serve as possible islands of reliability in the extraction of fundamental frequency. Section 8.1 addresses the automatic syllabification of an utterance from its phonetic realisation, and the syllabification of an utterance on the basis of a set of abstract phonological rules.

Microprosodic variations in the fundamental frequency of a speech waveform cause inflections in an *FØ* contour which are independent of the underlying intonational component. An *FØ* contour also includes speaker-dependent effects, fluctuations related to

cycle-to-cycle jitter (*FØ* perturbations) generated at the vocal folds, and erroneous *FØ* estimates generated by malfunctions in an FDA. It is proposed that an acoustic-phonetic representation of intonation can be derived from a raw *FØ* contour by removing components of the contour which are due to microprosody, cycle-to-cycle jitter, speaker dependent *FØ* range and fluctuations induced by erroneous *FØ* extraction from the speech signal. The task of reducing the occurrences of *FØ* perturbations and extraction errors is largely accomplished by the post-processing techniques described in Section 7.3.

The stylisation of an *FØ* contour aims to prevent microprosodic variations from being confused as pitch accents, and hence effectively isolate the microprosodic and intonational components. The process of piece-wise linear stylisation of an *FØ* contour is presented in Section 8.2.1. *FØ* estimates made within syllable nuclei are generally reliable because the speech signal within syllable nuclei is quasi-periodic and has a relatively high signal-to-noise ratio. These conditions are generally well suited to the assumptions made by *FØ* determination algorithms. Furthermore, microprosodic variations tend to dominate an *FØ* contour within the vicinity of short syllable nuclei in the context of unvoiced consonants. Syllable information can therefore be used to aid the stylisation of an *FØ* contour.

A stylised contour may contain some microprosodic variations given that the stylisation process will not be faultless. A stylised contour also contains speaker-dependent and intonational components. The stylised contour must be processed with respect to the syllables of an utterance in order to eliminate any remaining microprosodic variations, to compensate for speaker-dependent effects and to form an acoustic-phonetic representation of intonation. A schematisation process aims to manipulate a stylised contour with these goals in mind. A schematic representation of an *FØ* contour is independent of a speaker's fundamental frequency range, and aims to exhibit only the intonational component of the contour. An algorithm to generate *FØ* schemata is presented in Section 8.2.2. The *FØ* trajectories of a schema can be associated with syllables to locate the syllables which are pitch accented.

The processed acoustic parameters need to be combined to form a description of the intonation and the sentential stress patterns of speech, in a syllabic domain. The modular analyses proposed in this thesis are integrated to form a prosodic analysis system which

generates such a description. An overview of the system is described in Section 8.4 and its performance is evaluated in Section 8.5 relative to two algorithms formerly proposed in the literature.

8.1 Syllabification

The prosodic aspects of speech are described in a syllabic domain. Thus, the automatic identification of syllables in connected speech forms a part of prosodic analysis. An algorithm is proposed in Section 8.1.1 for the syllabification of speech from acoustic-phonetic parameters. The algorithm groups phones according to the phonetic realisation of an utterance rather than on the basis of a set of abstract phonological rules. A procedural definition of a syllable based on phonological rules is presented in Section 8.1.2. The groups of phones produced by the acoustic-phonetic syllabification are shown in Section 8.1.3 to correlate with syllables defined on the basis of phonological rules.

8.1.1 Automatic syllabification from acoustic-phonetic parameters

The following algorithm is used to group phones into syllable-sized units using a (frame-level) utterance-normalised low-band energy contour which is generated using the method described in Section 5.1, and the phone boundary and label information of an utterance provided either by hand or by an automatic segmentation system.

In the application of prosodic analysis for computer aided pronunciation teaching, the orthographic transcript of an utterance is known because a foreign language learner is asked to read a given sentence in the course material. An automatic segmentation algorithm can therefore use a network of variant pronunciations (which may also include optional interword silences and segmental assimilation) for the known utterance (McInnes *et al.*, 1992). The constraints of such a network enable automatic segmentation to be performed more reliably than in speech recognition systems where segmentation is either constrained only by vocabulary and syntax, or even worse, no vocabulary and syntax constraints are imposed (as in phoneme lattice generation).

The (auto-)segmentation data gives the location of phones and classifies phone-types. The low-band energy contour of an utterance is used to determine the degree of syllabic

cohesion between two adjacent phones.

All local minima (valleys) in the low-band energy contour are located. The locations of the minima form candidates for syllable boundaries. The areas of silence identified by the (auto-)segmentation are respected and the energy minima within such areas are believed to be due to variations in background noise. Each segmentation boundary at the junction of a silence and a phone label is regarded as a syllable boundary (at either the beginning or at the end of a syllable). The nearest candidate (energy minimum) to such a segmentation boundary is therefore moved to align with it and all the candidates residing within the area of silence are disregarded.

If word boundary information is known, then syllable boundaries can be forced to occur at word boundaries. The nearest syllable boundary candidate to a word boundary can be aligned with the word boundary. Word boundary information is not used in this study and so phones which are grouped using this algorithm may span across word boundaries. None of the analysis techniques described in Chapters 4, 5, 6 & 7 make use of word boundary information.

The regions between all the remaining energy minima (after those within silences have been disregarded) are taken to be potential syllables, with a start time given by the location of the nearest minimum on the left-hand side, and the location of the nearest minimum on the right-hand side giving the stop time. It is then determined whether or not the location of each of these potential syllables overlaps more than 50% of any vowel segment. If a potential syllable overlaps more than one vowel segment in this way, then the vowel segment with the maximum low-band energy is taken to be the nucleus of the syllable. If there is more than one vowel in a group of phones then any reduced vowels (/ə/) take a lower precedence in forming the syllable nucleus than non-reduced vowels, even if the reduced vowel has a greater low-band energy than the other vowel or vowels in the group of phones. (It is unlikely that a schwa will really have a higher low-band energy than a neighbouring full vowel, but this situation may arise because of errors in the automatic segmentation of an utterance.)

If no such overlap occurs, then it is determined whether or not the location of the potential syllable overlaps more than 50% of one of the possible syllabic consonant seg-

ments /l, m, n, (r)[1]/. Again, if the potential syllable overlaps more than one of these, the one with the maximum low-band energy is selected as the syllabic nucleus.

If there is insufficient overlap, the region between the minima does not correspond to a syllable unit (because no syllable nucleus can be found between the low-band energy minima) and either the left-hand side or right-hand side minimum is disregarded as a syllable boundary candidate — whichever has the highest energy and does not correspond to a phone/silence boundary or (if known) a word boundary. The newly formed region is then taken to be a potential syllable and the process is repeated.

The resultant syllabification has boundaries located in the utterance at positions of local minima in the low-band energy contour. These boundaries are aligned with the (auto-)segmentation by moving each of the boundaries to the nearest phone boundary.

8.1.2 Syllabification based on phonological rules

The following procedural definition is used to group phones into syllables on a phonological basis rather than on an acoustic-phonetic basis (Bagshaw & Williams, 1992).

i) Consonantal phones (such as /m, n, l, r, s/) which may result in schwa deletion (Gimson, 1970; Dalby, 1986) and take on the syllabic nucleus, are syllabified as if the underlying schwa is present. Hence, in rapid speech, *shortest* may be syllabified as ['ʃ ɔ – t s̩ t] and *additional* may be syllabified as [ə – 'd ɪ – ʃ n̩ – l̩]. A glottal stop /ʔ/ that may occur before or instead of a word-final stop consonant is treated as an instance of the underlying stop consonant, and the glottalised onset to a vowel is considered to be part of the vowel.

ii) Syllable boundaries are formed from the boundaries of words considered in isolation. Although in connected speech consonants at the end of one word can syllabify with the initial vowel of the following word (Maddieson, 1985), such resyllabification is not necessary in forming a domain in which to describe prosodic events. Thus, for example, the syllabification of *at all* differs to that of *a tall* even if [t] is aspirated in both cases and they are phonetically identical. Similarly, resyllabification is unnecessary across words that appear to blend together due to vowel deletion, as may be the case in *under a*, which is syllabified as ['ʌ n – d r̩ – ə] rather than ['ʌ n – d r ə]. The syllabification of

[1] /r/ is included for American English and for rhotic dialects of British English.

under a and *tundra* /'t ʌ n – d r ə/ therefore differ even though they may be phonetically similar. This approach is adopted because the exact boundaries between syllable nuclei are not of critical importance, although identifying the syllable nuclei is important.

iii) The boundaries between syllables are also determined by the presence of morphological boundaries. The boundary between a free morpheme and an inflectional suffix (except *-s* — eg. *-tion*, *-tician*) or a class-II derivational affix (eg. *un-*, *non-*, *dis-*) is taken to be a syllable boundary (see Fudge (1984) for details of morpheme prefixes and suffixes in English). Thus, *hopeless* is syllabified as /'h əʊ p – l ə s/ rather than /'h əʊ – p l ə s/; and *uninteresting* is syllabified as /ʌ n – 'ɪ n – t ə – r ɛ s t – ɪ ŋ/ rather than /ʌ – 'n ɪ n – t ə – r ɛ – s t ɪ ŋ/.

iv) On the basis of English phonotactics, any cluster of phones forming the onset or the coda of a syllable must also be a permissible word-initial or word-final cluster. According to this rule, *extra* may be syllabified as /'ɛ k – s t r ə/, /'ɛ k s – t r ə/, or /'ɛ k s t – r ə/.

v) The 'maximal onset (and minimal coda) principle' (Pulgram, 1970; Couper-Kuhlen, 1986) arbitrates between competing syllabifications. According to the principle, as many consonantal phones as possible form a syllable onset. Using this principle, *extra* would be syllabified as /'ɛ k – s t r ə/. However, in cases when alternative boundaries are possible, stressed syllables tend to attract consonants more than unstressed ones, particularly in the case of ambisyllabic consonants such as /s, f/ (Fudge, 1984). When this final criterion is applied, the syllabification adopted for *extra* becomes /'ɛ k s – t r ə/.

8.1.3 Evaluation procedure

An example of the syllabifications produced by the above methods is shown in Figure 8.1 for the word *international*, in the phrase "...at the Kyoto International Conference Centre." The upper part of the Figure shows the speech waveform and its corresponding segmental transcription (obtained by auto-segmentation). Phone boundaries are shown by dotted lines and the continuous lines show the syllable boundaries derived on a phonological basis. The lower part gives the utterance-normalised low-band energy contour and transcription-aligned syllable boundaries derived using the acoustic-phonetic syllabification algorithm described in Section 8.1.1. The phonologically based syllabification

gives five syllables, /ɪ n – t ɚ – n æ – ʃ ə n – ə l/, but the acoustic-phonetic syllabification gives only four syllables, [ɪ n – t ɚ – n æ ʃ – ə n ə l]. The final syllable [– ə n ə l] may initially appear to illustrate an error in the acoustic-phonetic syllabification algorithm (that of a *missing* syllable boundary). However, the low-band energy contour shows only one peak of intensity in this section of speech, and it is transcribed phonetically by a phonetician as [– n ə l] (with schwa deletion). There is an error in the automatic segmentation of the utterance, but not in its subsequent syllabification.

A database of 591 utterances from the English language ATR conference-registration dialogues is syllabified automatically using the above acoustic-phonetic algorithm and by hand using the syllabification based on phonological rules. The data consists of a series of telephone dialogues spoken by a female bilingual speaker of American English and Japanese in a variety of styles and registers, ranging from read speech to spontaneous speech (see Campbell (1992) for a brief description of this material).

In order to compare the human and machine syllabifications, it is necessary to define when they are regarded as *sufficiently similar* and when they are not. At the most stringent level of comparison, a syllable located automatically is regarded as correct only if its boundaries (both at the beginning and at the end of the syllable) match exactly with those defined on a phonological basis. This method of comparison is unrepresentative of the algorithm's performance as just one miss-matched boundary would correspond to two miss-matched syllables. Furthermore, the boundaries between syllables vary even between phonological definitions — only, the syllable nuclei are well defined. Therefore, the following procedure is used to evaluate the automatic syllabification algorithm.

Initially assume that all of the syllables located automatically are *extra* syllables — ie. that each auto-syllable is not the single match of a phonological syllable. Also assume that all of the phonological syllables are *missing* in the automatic syllabification — ie. that each hand-syllable is not matched by any syllables located automatically. The phonetic transcription of every syllable is known. By definition, there is at least one vowel and/or syllabic consonant in each syllable (defined either automatically or on a phonological basis). However, the phonetic units which form each syllable nucleus are not stipulated (as they may differ from one syllabification scheme to another). The comparison proceeds by considering each hand-syllable in turn. Locate the first vowel

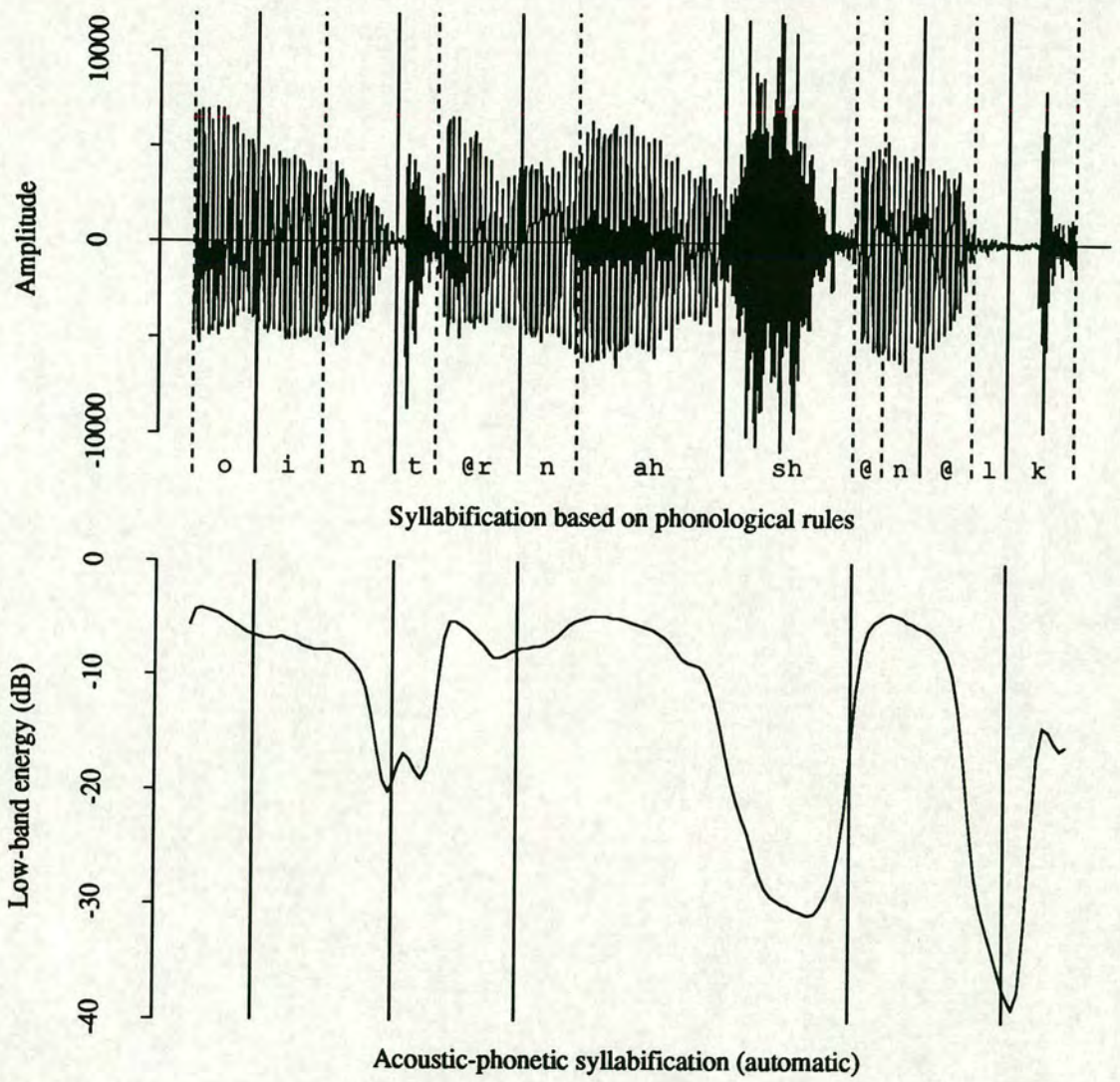


Figure 8.1: Syllabification of the word *international*

or syllabic consonant (a potential nucleus) in the hand-syllable. Then consider each auto-syllable from the beginning of the utterance to determine whether or not it overlaps with the potential nucleus. If the auto-syllable does overlap and if it has previously been assumed to be an *extra* syllable, then mark the auto-syllable as not being an *extra* syllable, mark the hand-syllable as not *missing*, and consider the next hand-syllable in the utterance. Otherwise, repeat the search for a matching auto-syllable using another

Total number of syllables from a phonological basis (by hand)	from an acoustic-phonetic basis (automatic)	Match	Missing	Extra
9322 (100.0%)	8910 (95.6%)	8875 (95.2%)	-447 (4.8%)	+35 (0.4%)

Table 8.1: Comparison of syllabifications based on phonological rules (by hand) and on acoustic-phonetic parameters (automatic)

potential nucleus in the hand-syllable until all possible nuclei have been considered. If none of the possible nuclei in a hand-syllable are overlapped by an auto-syllable previously assumed to be an *extra* syllable, then the hand-syllable is genuinely *missing* in the automatic syllabification. Once every hand-syllable in the utterance has been considered in turn, any remaining auto-syllables which have not been found to correspond with a hand-syllable are genuinely *extra* syllables.

There is a large level of agreement between the syllabic domains generated automatically and by the procedural definition based on phonological rules (see Table 8.1). The *missing* syllable boundaries are due to the occurrences of vowel/vowel boundaries between which there is no valley in the low-band energy. When this case arises, often one of the vowels is a schwa; for example, the phonological syllabification of *my address* as /m aɪ – ə – d r ɛ s/ can be grouped on an acoustic-phonetic basis as [m aɪ ə – d r ɛ s]. Conversely, *extra* syllable boundaries occur when the low-band energy dips within the phonologically based syllable at a vowel/vowel boundary or vowel/syllabic consonant boundary; for example, the phones in *tour* /t ʊ ɚ/ can be grouped as [t ʊ – ɚ] on an acoustic-phonetic basis, and the phones in the word *forms* /f ɔ r m s/ are grouped as [f ɔ – r m̩ s] when its phonetic realisation tends towards the pronunciation of *forums* with schwa deletion, but with a fall in low-band energy remaining between /ɔ/ and /m/.

The algorithm described in Section 8.1.1 forms groups of phones with a vowel or syllabic consonant as the nucleus of each group. There are no vowel/vowel pairs or vowel/syllabic consonant pairs with dips in low-band energy between them, in any of the groups. The boundaries between groups are positioned at the point of minimum low-band energy between nuclei (aligned to the nearest phone boundary). It is assumed that such groups of phones can only ever be perceived as single prominent units in

connected speech. These units are shown to correlated closely with syllables defined using phonologically based rules. They are, therefore, referred to as *syllables*.

8.2 Processing of *FØ* for an acoustic-phonetic description of intonation

The fundamental frequency contour of a speech waveform does not form an acoustic-phonetic representation of the utterance intonation because microprosodic variations are also present and the contour is dependent upon a speaker's fundamental frequency range. The process of piece-wise linear stylisation of a contour presented in Section 8.2.1 aims to prevent microprosodic variations from being confused as pitch accents. A stylised contour is processed with respect to the syllables of an utterance in Section 8.2.2 to form a schematic acoustic-phonetic representation of intonation. The schematic representation is independent of a speaker's fundamental frequency range.

8.2.1 Linear piece-wise stylisation of an *FØ* contour

The algorithm used to perform the stylisation of an *FØ* contour is based on the technique proposed by Scheffers (1988). As with other *FØ* contour stylisation techniques (Hirst & Espesser, 1993; Taylor, 1993), the *FØ* contour is the only input parameter to Scheffers' algorithm. Information about the segmental content of an utterance is not included. The algorithm described here to stylise an *FØ* contour also uses syllable nucleus boundaries (which are automatically located using the acoustic-phonetic syllabification algorithm described in Section 8.1.1).

Scheffers partitions an *FØ* contour into linear piece-wise sections using the least (mean of) squares method of regression analysis. The least (mean of) squares method of regression analysis produces its best results when the underlying error distribution is Gaussian. This method of regression analysis becomes unreliable if the data contains spurious *outliers* (samples which deviate from the local trend by a large amount). Such data almost invariably exists amongst *FØ* contours, where gross measurement errors and microprosodic variations can, and often do, occur. A more robust method of regression analysis is therefore required. The least median of squares (LMedS) method of regression

analysis (Appendix A) is more robust to the presence of outliers (Rousseeuw & Leroy, 1987). The algorithm described here to stylise an *FØ* contour incorporates the robust least median of squared residuals regression analysis.

An account is given below which describes how *significant* turning-points are located in sections of an *FØ* contour (partitioned using syllable nucleus boundaries) then modified to prevent contour discontinuities other than at the boundaries between unvoiced and voiced speech, and how a new, stylised contour is generated by interpolating between the turning-points.

The *FØ* values describing a contour (excluding values which equal zero to represent unvoiced speech) are initially converted to the semitone scale using Equation 8.1.

$$FØ_{semitone} = 12.0 \log_2 \left(\frac{FØ_{Hz}}{55.0} \right) \tag{8.1}$$

An *FØ* contour is partitioned into sections which represent portions of continuously voiced speech and which overlap any part of at least one syllable nucleus. Partitioning an *FØ* contour in this way eliminates isolated voiced sections of the contour which do not overlap a syllable nucleus from the subsequent processing (which locates significant turning-points in the contour). This is required because isolated voiced sections in the *FØ* contour which do not overlap a syllable nucleus are likely to be unreliable since they may correspond to unvoiced sections of speech erroneously classified as voiced by an FDA.

Short sections (less than 40ms in length) of an *FØ* contour partitioned as above correspond to short syllable nuclei with unvoiced left and right segmental contexts. Thus, microprosodic variations at the onset and offset of voicing dominate the *FØ* trajectory in such short sections. Significant turning-points are therefore not located in these short sections of a contour. They are stylised by setting the *FØ* values to the mean value of the *FØ* trajectory within each short section.

The following process (adapted from Scheffers, 1988) is used to identify the turning-points in each section of an *FØ* contour which is longer than 40ms. Starting with the first voiced frame, LMedS regression analysis is applied to a window of w frames corresponding to voiced speech, where w is initially set to 5. The final frame in this window is taken to be a turning-point candidate. The *FØ* value of the subsequent frame is predicted using

the coefficients of the LMedS regression analysis. If the absolute difference between the actual and predicted *FØ* values is less than or equal to some level of permitted variation in *FØ* (1 semitone), then the candidate is not a turning-point, the window length w is incremented to include the next voiced frame, and the above process is repeated. The repetition of this process terminates when the turning-point candidate is the final voiced frame in the section of the *FØ* contour under analysis. Otherwise, when the absolute difference is greater than the permitted *FØ* variation, either this subsequent *FØ* value constitutes some type of irregularity in the *FØ* contour or the candidate could be a genuine turning-point. In order to determine which is the case, the *FØ* value of the next voiced frame is also predicted. If the absolute difference between the predicted value and the actual value is once again greater than the permitted *FØ* variation, and if this situation arises for all following frames up to either the final voiced frame in the section of the *FØ* contour under analysis or such that the duration of this discontinuity is greater than some minimum permitted level (100ms), which ever occurs first, then the candidate is said to be a genuine turning-point. Otherwise, the length of the window w is increased to include the first frame for which the absolute difference in the actual and the predicted *FØ* values was less than or equal to the permitted variation, but not to include those for which it was greater than the permitted variation, and the LMedS regression analysis process is repeated. If the candidate was found to be a turning-point and if it corresponds to a voiced frame immediately preceding a frame of unvoiced speech, then the first frame of the next voiced region is also designated as a turning point. This entire process is then repeated with the length of the window w reset to 5 and with the first frame of the window set to the frame of the most recent turning-point found. The first and final voiced frames of the non-stylised contour are also assigned as turning-points.

In order to ensure that discontinuities in the stylised *FØ* contour only occur at unvoiced sections of speech, the fundamental frequency at each turning-point of the new contour is determined in a way which depends upon the voicing state of the frames adjacent to it. However, a discontinuity in the stylised *FØ* contour is allowed within a voiced section of speech if the turning-point is an *outlier* for either of the piece-wise sections it joins. (The detection of outliers is a part of the LMedS regression analysis.) In such situations, the turning-point is treated as being adjacent to an unvoiced frame of

speech. For any given turning-point (tp) at frame f_{tp} with original fundamental frequency $F\emptyset_{tp}$, the LMedS coefficients s_{tp} (slope) and i_{tp} (intercept) of selected points preceding the turning-point are known. The modified fundamental frequency $F\emptyset'_{tp}$ is calculated from Equation 8.2.

$$F\emptyset'_{tp} = \begin{cases} 0.5(s_{tp}.f_{tp} + i_{tp} + s_{tp+1}.f_{tp} + i_{tp+1}) & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ are voiced} \\ s_{tp+1}.f_{tp} + i_{tp+1} & \text{if frame } f_{tp}-1 \text{ is unvoiced \& frame } f_{tp}+1 \text{ is voiced} \\ s_{tp}.f_{tp} + i_{tp} & \text{if frame } f_{tp}-1 \text{ is voiced \& frame } f_{tp}+1 \text{ is unvoiced} \\ F\emptyset_{tp} & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ are unvoiced} \end{cases} \quad (8.2)$$

The new stylised contour is then created by linear interpolation of $F\emptyset$ between each turning-point (f_{tp}, $F\emptyset'_{tp}$) and by resetting each frame that is unvoiced in the non-stylised contour to an unvoiced state in the new contour. The resultant data is then coverted back to a Hertz scale.

An example of a linear piece-wise stylised contour produced using the above algorithm is illustrated in Figure 8.2. The stylised $F\emptyset$ contour is aligned with the speech waveform and phonetic transcription. The dotted lines across the speech waveform represent phone boundaries and the solid lines represent syllable boundaries, which are used by the stylisation algorithm. Many of the microprosodic $F\emptyset$ variations and $F\emptyset$ perturbations are removed by the stylisation process, however, some remain in sections of the contour associated with voiced consonants in a vocalic context. The overall melody of the utterance is retained.

8.2.2 Prosodic schematisation of a stylised $F\emptyset$ contour

Each piece-wise section in the stylised $F\emptyset$ contour forms a possible pitch accent or a possible part of an accent. Some piece-wise sections do not correspond to part of any pitch accent, such as those which are a direct consequence of erroneous $F\emptyset$ extraction or stylisation. Thus, a piece-wise section is treated as being part of a possible pitch accent if either at least 50% of the piece-wise section overlaps a syllable nucleus (where $F\emptyset$ estimation is expected to be reliable) or at least 50% of a syllable nucleus is over-

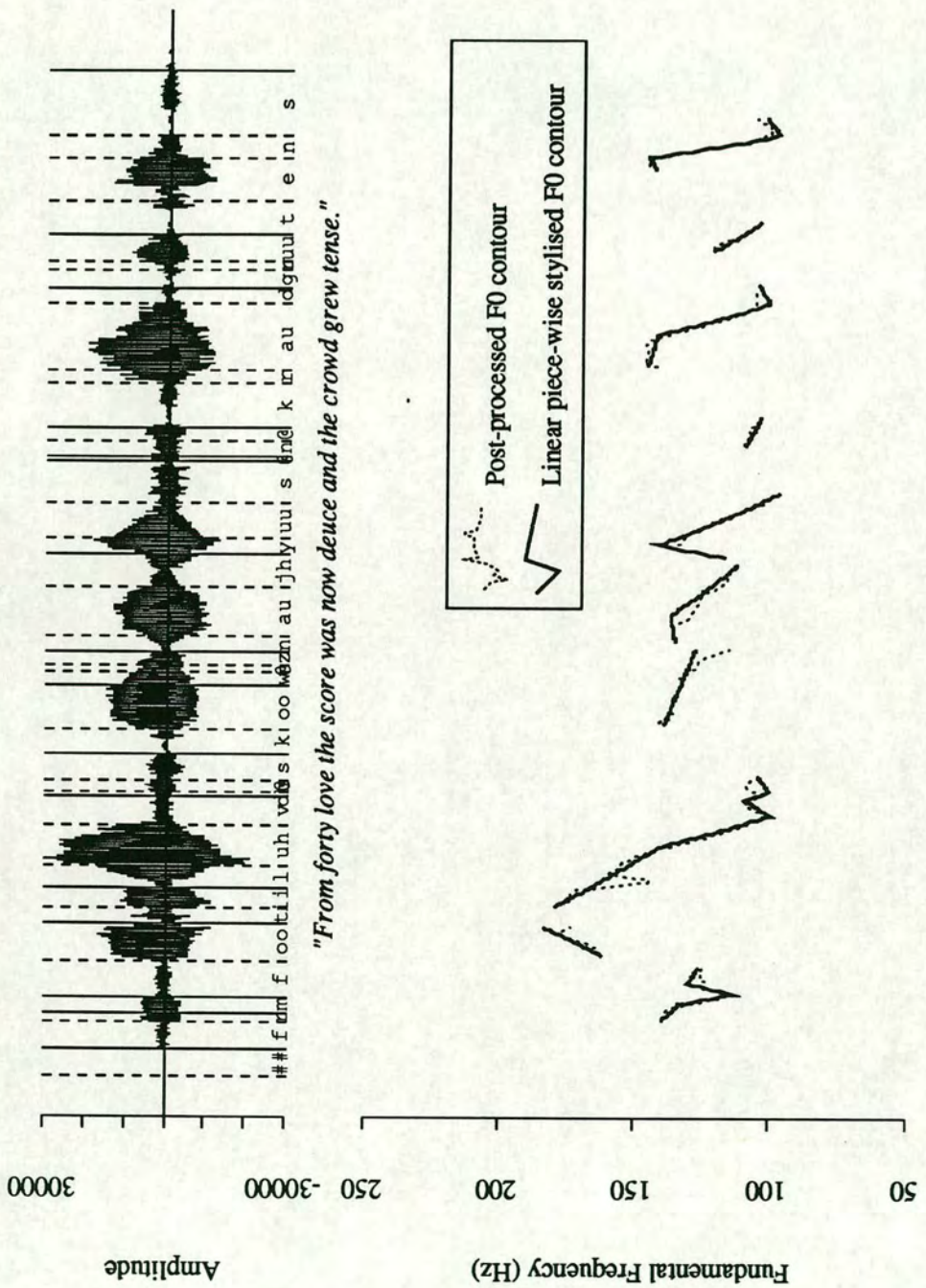


Figure 8.2: Linear piece-wise stylisation of an *FØ* contour*

lapped by the piece-wise section. The piece-wise sections may therefore extend beyond a syllable nucleus but only those crossing a syllable nucleus by a substantial amount are selected. This approach compromises between using information about the trajectory of *FØ* through vowels alone (which may be limiting for short nuclei) and using the *FØ* contour of an entire syllable (where *FØ* discontinuity errors may occur).

The absolute *FØ* range in an utterance will vary from speaker to speaker and from utterance to utterance (Section 3.4). *FØ* piece-wise sections are, therefore, normalised for each utterance to give relative *FØ* heights. The relative height of each piece-wise section is calculated by first locating a regression line which best fits the contour turning-points using LMedS analysis. A by-product of the LMedS is the standard deviation σ_{LMedS} of the points from the resultant linear model. The absolute *FØ* at each turning-point is then converted by subtracting its modelled value and dividing by the standard deviation, σ_{LMedS}. This effectively compensates for any long term declinative tendency that may be exhibited in the fundamental frequency contour and expresses the *FØ* values relative to an utterance dependent datum.

Once the relative height of each piece-wise section has been established, the piece-wise sections are combined to form *FØ* trajectory descriptors. The *FØ* trajectory descriptors used are level, fall, rise, fall-rise and rise-fall "–, \, /, ∨, ∧". Each piece-wise section which crosses a substantial part of a syllable nucleus (as described above) is classified as either level, a fall, or a rise. If there is no piece-wise section which crosses a particular syllable nucleus by a substantial amount, then the syllable is classified as having an "unknown" *FØ* trajectory. Let $F\emptyset_{start}$ represent the relative *FØ* height at the start of the piece-wise section and that at the end of the section be represented by $F\emptyset_{end}$. A piece-wise section is classified on the basis of Equation 8.3.

$$F\emptyset \text{ trajectory} = \begin{cases} \backslash & \text{if } F\emptyset_{start} - F\emptyset_{end} > 0.75\sigma_{LMedS} \\ / & \text{if } F\emptyset_{start} - F\emptyset_{end} < -0.75\sigma_{LMedS} \\ - & \text{otherwise} \end{cases} \tag{8.3}$$

When two or more piece-wise sections cross any particular nucleus, they are combined by initially taking all adjacent sections with the same *FØ* trajectory descriptor and

joining them into one. A join is made by setting $F\emptyset_{start}$ to that of the first section, $F\emptyset_{end}$ to that of the second section, and reclassifying using Equation 8.3. In the combined training data and test data of 660 utterances (described in Section 4.1) consisting of 8546 syllables, there are only six syllables for which more than two piece-wise sections remain after this process. One of these six syllables contains a section of $F\emptyset$ halving errors, two syllables contain extreme $F\emptyset$ perturbations due to creaky voice, and three syllables include microprosodic variations which are not successfully removed by the stylisation process. If there are two remaining sections (their classifications must differ) and if either is classified as level "–", then they too are joined in the same way. Otherwise, one section is a fall "\" and the other section is a rise "/". These sections are combined to give a single trajectory which is described as either a fall-rise "V" or a rise-fall "Λ" depending on their order, and the relative level at their mid-point is kept for reference. Thus, for the fall-rise and rise-fall descriptors, the relative heights of the onset, the mid-point and the offset of the trajectory are known.

$F\emptyset$ schemata are generated from sequencies of $F\emptyset$ trajectory descriptors and their internalised relative heights. An example of the processing of a speech waveform to generate an $F\emptyset$ schema is illustrated in Figure 8.3. The speech waveform for the sentence, "From forty love the score was now deuce and the crowd grew tense," is shown with its aligned phonetic transcription represented by MRPA symbols (Appendix D). The raw $F\emptyset$ contour is extracted using the super resolution $F\emptyset$ determinator (eSRFD) and includes $F\emptyset$ doubling and halving errors. These errors are eliminated by the de-step filter and non-linear smoothing algorithm (Section 7.3). The linear piece-wise stylisation of the post-processed $F\emptyset$ contour removes many of the microprosodic $F\emptyset$ variations. The $F\emptyset$ schema derived from this exhibits only the intonational component of a raw $F\emptyset$ contour.

Judgements of syllable accentuation are based on the $F\emptyset$ trajectory descriptors using the pitch accent decision filter developed by Hieronymus (1989) — illustrated in Figure 3.5. The decision filter examines three $F\emptyset$ trajectory descriptors for each syllable — the first syllable $F\emptyset$ trajectory descriptor which is not classified as "unknown" on the left-hand side of the syllable, the $F\emptyset$ trajectory descriptor of the syllable, and the first syllable $F\emptyset$ trajectory descriptor which is not classifier as "unknown" on the right-hand

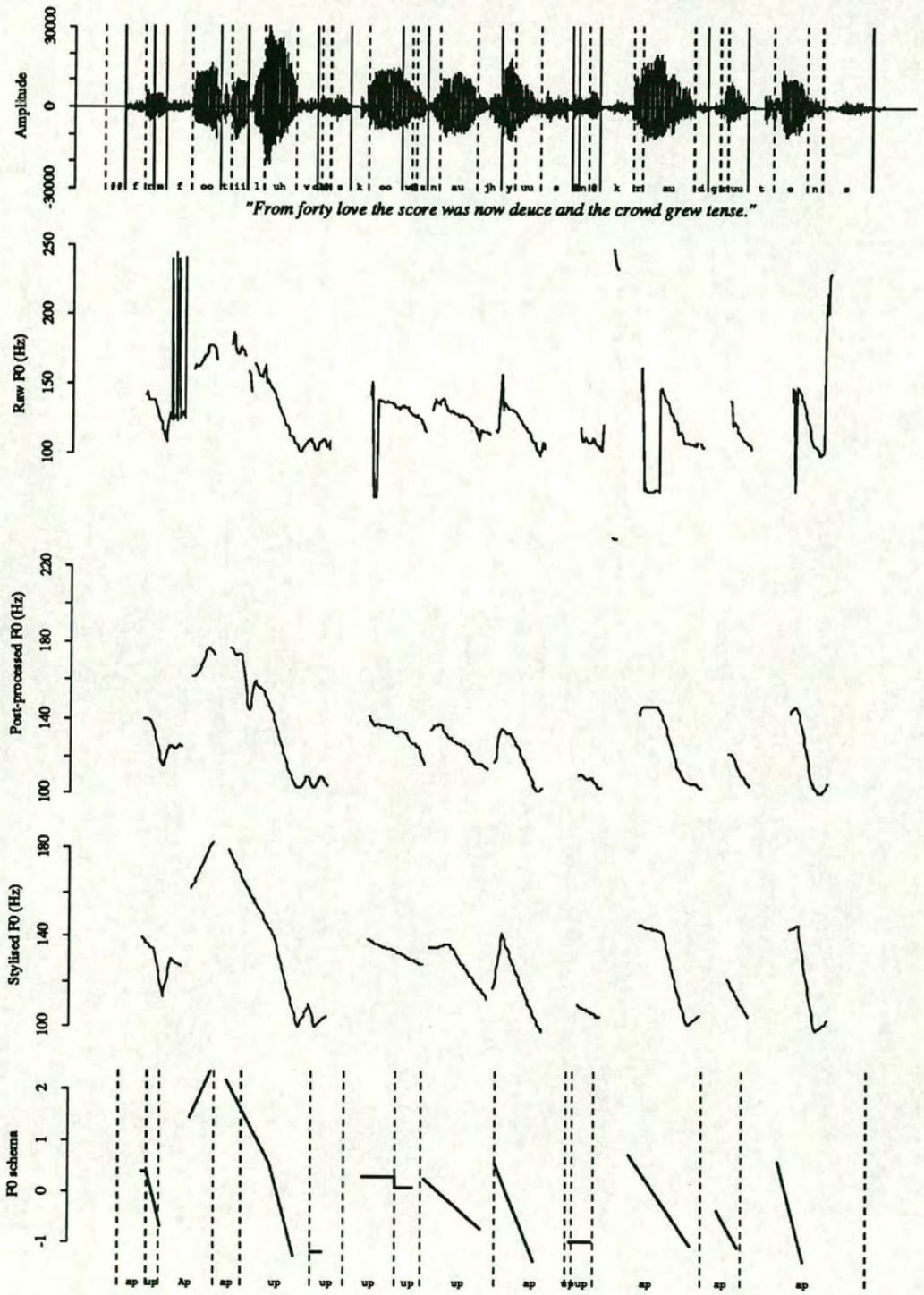


Figure 8.3: Transformation from a raw *FØ* contour to a schematic acoustic-phonetic representation of intonation*

side of the syllable. If the *FØ* trajectory of a syllable is either a fall-rise "V" or a rise-fall "Λ" then the syllable is classified as being accented. Syllables with an "unknown" *FØ* trajectory are classified as being unaccented. If the *FØ* trajectory descriptor of the left-hand syllable context or the right-hand syllable context is either a fall-rise "V" or a rise-fall "Λ", then only the half of the fall-rise or rise-fall closest to the syllable is used by the decision filter.

Syllables classified as accented by the decision filter are labelled as "*ap*". A syllable is also classified as accented if it corresponds with the maximum relative *FØ* height of an *FØ* schema. Syllables corresponding to such maxima are labelled "*Ap*". All other syllables are labelled "*up*". The *FØ* schema in Figure 8.3 is shown with syllable boundaries and the syllable accentuation labels produced by this method.

The levels of agreement between the categories of syllable accentuation as transcribed by hand and as assigned automatically are shown in Table 8.2. Syllables which are transcribed by hand as stressed but unaccented "*s*" or unstressed "*u*", are collectively referred to as unaccented syllables "*UA*". The syllables which are transcribed by hand as pitch accented "*PA*" do not include non-stress accents. (Refer to Section 2.2.4 for a clarification of these terms.) The pitch accent decision filter does not include the location of prominent syllables as an input parameter. Syllables transcribed as accented "*ap*" by the decision filter therefore include non-stress accents. This accounts for the large number of unaccented syllables "*UA*" which are confused as pitch accented "*ap*" by the decision filter.

8.3 Combination of acoustic parameters

The unidimensional contribution to the perceived prominence of a syllable made by features of duration and energy is investigated in Chapters 4 & 5. Sections 4.5 & 5.4 show that the duration feature $D_{feature}$ (Equation 4.6) and the energy feature $E_{feature}$ (Equation 5.3) can be used in conjunction with a simple peak-picking algorithm to label syllables as either prominent or non-prominence and respectively yield a 77.8% and an 84.0% agreement level with prominence labels based on human perception (in an open test).

Section 8.2.2 shows that the trajectories of an *FØ* schema can be used in conjunction

		Pitch accent algorithm label			
		Ap	*ap*	*up*	total
Hand Label	*PA*	359 (6.7%)	992 (18.7%)	486 (9.1%)	1837 (34.5%)
	UA	101 (1.9%)	864 (16.2%)	2517 (47.3%)	3482 (65.5%)
	total	460 (8.6%)	1856 (32.0%)	3003 (56.5%)	5319 (100.0%)

Correct classification rate = 3868/5319 (72.7%)

Table 8.2: *FØ* schema: Confusion matrix for pitch accent decision filter (*Ap* — syllable with maximum *FØ* in schema; *ap* — accented syllable according to decision filter; *up* — unaccented syllable according to decision filter; *PA* — pitch accented syllable; *UA* — unaccented syllable)

with a pitch accent decision filter to label syllables as either accented or unaccented and yield a 72.7% agreement level with syllable accentuation labels based on human perception. The accented syllables identified by the pitch accent decision filter include non-stress accents.

Syllable classification labels are generated by the peak-picking algorithm for the duration and energy features, and by the pitch accent decision filter for the trajectories of the *FØ* schema. These labels are combined to classify each syllable in an utterance as either pitch accented "*PA*", stressed but unaccented "*s*", or unstressed "*u*". These labels are assigned using a procedure adapted from the equal-weight two-out-of-three voting scheme proposed by Hieronymus (1989).

A syllable is categorised as prominent if it is associated with an "*Sd*" label (maximum $D_{feature}$), an "*Se*" label (maximum $E_{feature}$) or an "*Ap*" label (maximum in *FØ* schema). A syllable is also categories as prominent if it is associated with at least two-out-of-three of the labels "*sd*" (local maximum in $D_{feature}$ contour), "*se*" (local maximum in $E_{feature}$ contour) and "*ap*" (accented syllable according to the pitch accent decision filter). A syllable which is categorised as prominent by these rules is labelled as being pitch accented "*PA*" if it is associated with either an "*Ap*" or an "*ap*" label; otherwise it is labelled as being stressed but unaccented "*s*". All other syllables are categorised as unstressed and are labelled "*u*".

Figure 8.4 shows the confusions between perceived prominence labels (transcribed

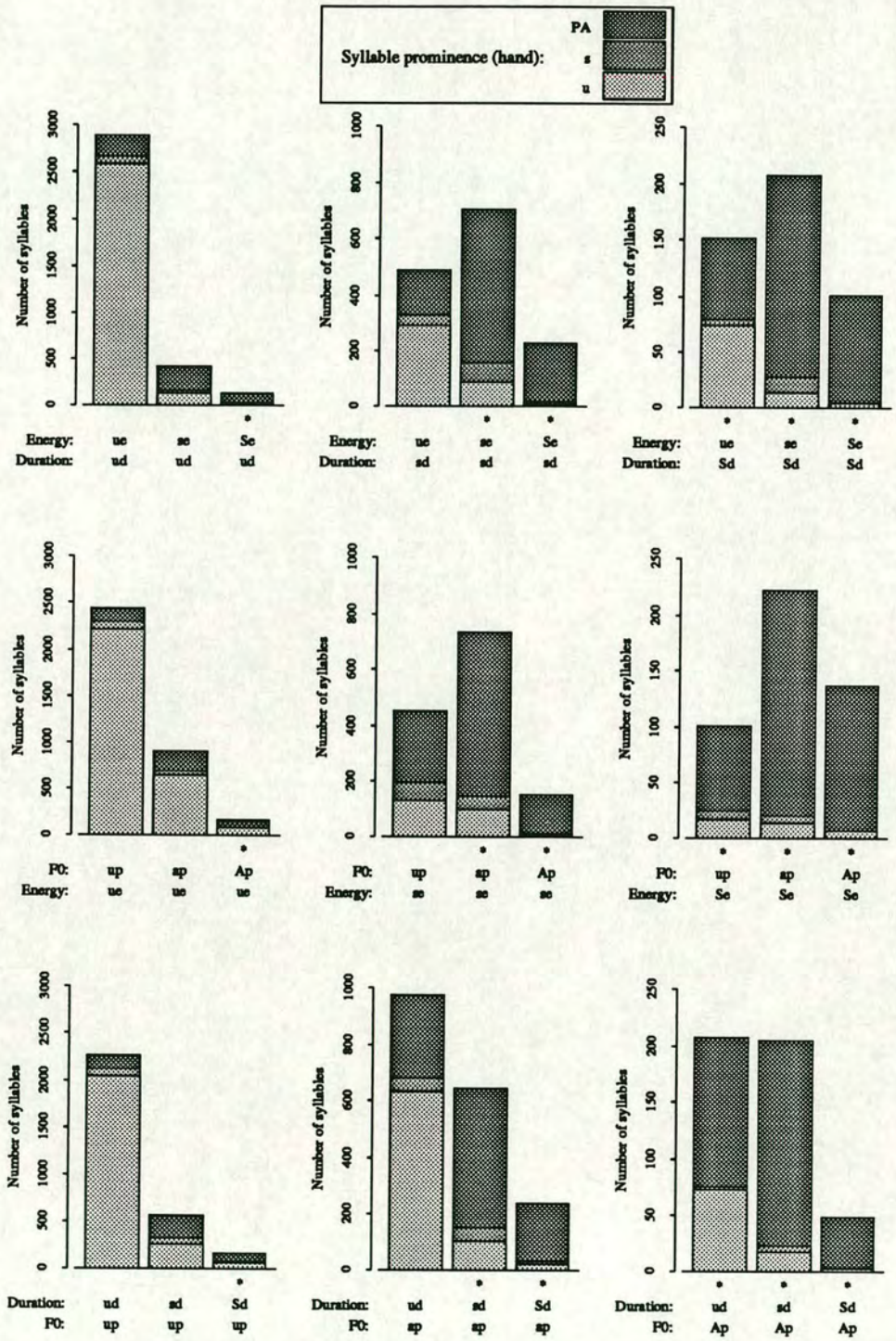


Figure 8.4: Multi-way confusions across acoustic parameters

by hand) and each possible pair of labels generated by the processing of the acoustic parameters. The combinations of labels which are marked by an asterisk are those which result in a syllable being labelled as prominent (both accented and unaccented) by the procedure described above. There are more prominent syllables than non-prominent syllables (transcribed by hand) associated with the combinations of automatically assigned labels which are marked by asterisks.

8.4 Overview of the integrated prosodic analysis system

The prosodic analysis of speech described in this thesis involves a number of intergated modules. An overview of the automatic prosodic analysis system is shown in Figure 8.5. There are three underlying channels of analysis — one channel for each acoustic parameter (energy, duration and fundamental frequency). A direct flow of information through these channels from the input speech signal to the combination procedure is interrupted by the syllabification module. The syllabic domain ties together the information embedded within the three acoustic parameters. The syllabification module incorporates information about the low-band energy contour and the segmentation of an utterance. This information is passed on in an abstracted form to the normalisation of energy and duration measures and to the processing of an *FØ* contour.

A summary of the modules which make up this system to form a procedural analysis of an utterance is as follows.

Front-end signal processing:

- Transform the speech signal from the time domain to a frequency domain representation by applying an FFT to 20ms frames of data at 5ms intervals.
- Extract the (frame-level) low-band energy contour from the frequency domain representation of the speech signal.
- Cepstral coefficients are required for the automatic segmentation of speech into acoustic-phonetic units. The cepstral analysis is performed on 20ms frames of data at 5ms intervals.
- Segment the speech waveform into acoustic-phonetic units either by hand (for the

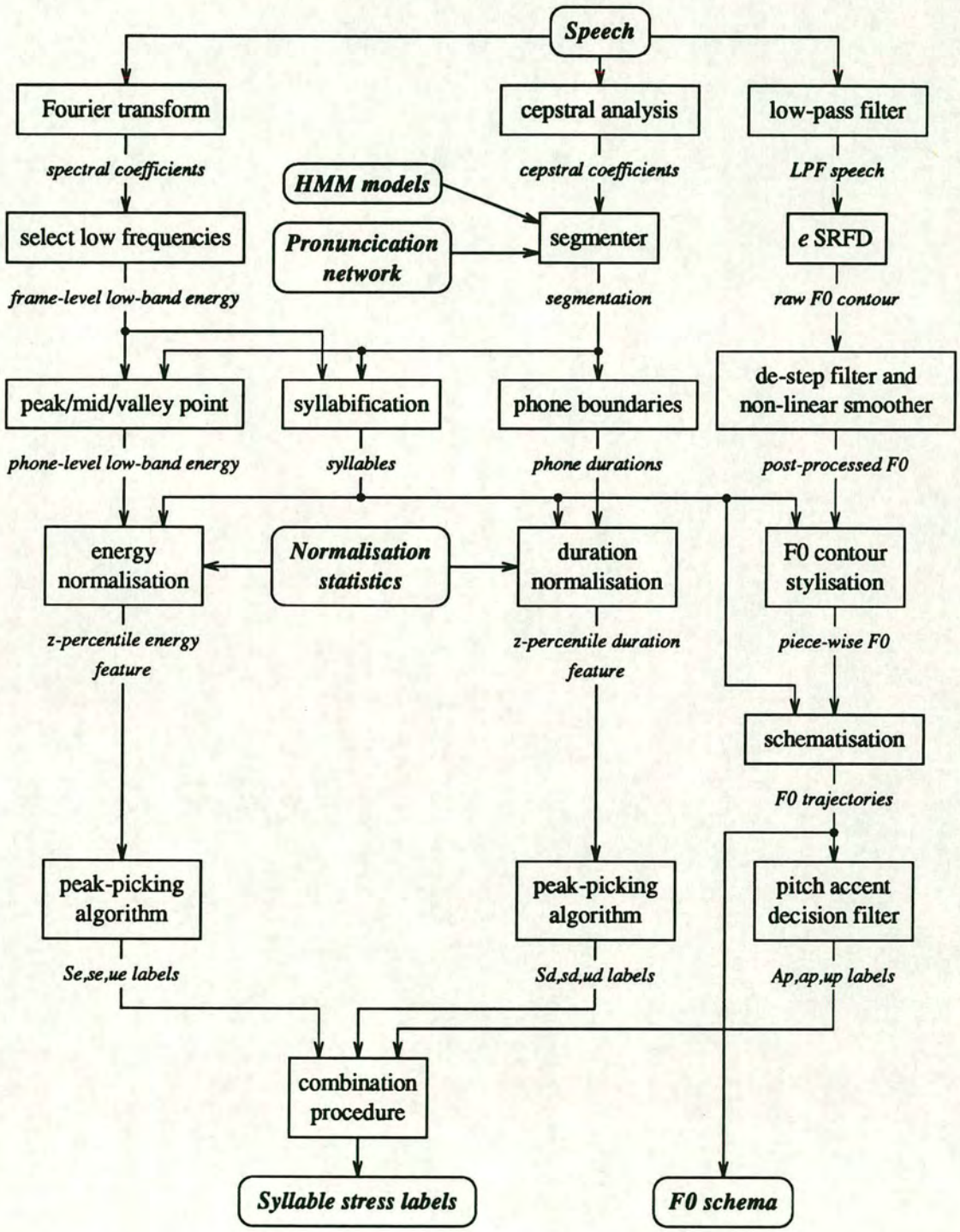


Figure 8.5: Block diagram of integrated prosodic analysis system

training data) or by an automatic segmentation system (McInnes *et al.*, 1992). This provides the phone boundary and label information of an utterance.

- Low-pass filter the speech signal using a finite impulse response (FIR) filter with a -3dB cut-off at 600Hz and rejection greater than -85dB above 700Hz.
- Extract the *FØ* contour from the low-pass filtered speech waveform using the enhanced super resolution *FØ* determination (*e*SRFD) algorithm. *FØ* values are generated at 5ms intervals.
- Apply the de-step filter and non-linear smoothing to the raw *FØ* contour in order to eliminate *FØ* doubling and halving errors and reduce *FØ* cycle-to-cycle perturbations.

Phone-level abstraction of acoustic parameters:

- Calculate the phone-level low-band energy contour from the frame-level low-band energy contour and the segmentation. A peak/valley/mid-point is selected in the energy contour for each phone.
- Obtain an acoustic-phonetic syllabification of the speech from the frame-level low-band energy contour and the segmentation. The syllabification identifies syllable nuclei and provides information about phone grouping.
- Absolute phone durations are obtained from the distance between phone boundaries in the segmentation.

Syllable-level abstraction:

- Calculate the normalised phone-level low-band energy feature $E_{feature}$ (Equation 5.3) for each syllable nucleus. The variations in $E_{feature}$ from syllable to syllable form a normalised energy contour.
- Calculate the normalised phone duration feature $D_{feature}$ (Equation 4.6) for each syllable lhyme. The variations in $D_{feature}$ from syllable to syllable form a normalised duration contour.

- Stylise the post-processed *FØ* contour into a sequence of linear piece-wise sections. The stylisation process makes use of syllable nucleus boundaries and reduces microprosodic variations in the contour.

Sentential stress and intonation representations:

- Schematise the piece-wise stylised *FØ* contour to identify *FØ* trajectories related to pitch accents. The *FØ* schema provides an acoustic-phonetic representation of the utterance intonation.
- Apply the pitch accent decision filter to the *FØ* schema in order to locate accented syllables.
- Apply the peak-picking algorithm to the normalised energy contour. This represents the unidimensional contribution made by energy to the prominence of a syllable.
- Apply the peak-picking algorithm to the normalised duration contour. This represents the unidimensional contribution made by duration to the prominence of a syllable.
- Finally, combine the syllable accentuation, energy and duration labels to categorise each syllable as either pitch accented "*PA*", stressed but unaccented "*s*", or unstressed "*u*".

8.5 Performance evaluation of prosodic analysis algorithms

The syllable prominence labels in the training and test data described in Section 4.1 are transcribed on the basis of human perception. This transcription, however, cannot be regarded as definitive. In order to give an objective evaluation of the integrated prosodic analysis system proposed in this thesis (referred to as the *PCB-system*), the performance of the system is compared with that of two algorithms formerly proposed in the literature, relative to the transcription based on human perception. The two former algorithms which are evaluated, are the algorithm based on a Bayesian classifier (Waibel, 1988) (referred to as the *AW-system*) and the algorithm which uses knowledge-based rules to automatically transcribe syllable prominence levels in connected speech

(Hieronymus, 1989; Hieronymus & Williams, 1991) (referred to as the *JLH-system*). The AW-system and the JLH-system are reviewed in Section 3.5.

Although the AW-system and the JLH-system have been evaluated by their respective developers, it is not possible to draw comparisons of their reported performances. Each evaluation depends on the form of the prosodic transcription which is taken as a reference and on the nature of the test speech and the data used to train the algorithms. In the evaluations presented here, the algorithms are both trained and tested using the data described in Section 4.1. A direct comparison can therefore be made between the performance evaluations of the prosodic analysis system proposed in this thesis and of the two former analysis techniques. The statistical significance of the differences in performance is determined.

8.5.1 Integrated prosodic analysis (PCB-system)

The normalisation statistics used to calculate the duration and energy features are determined from the training data. The statistics are determined from two inspections of the training data.

The first inspection determines the percentage p of non-prominent phones, the distributions of phone durations in which phones are grouped on a broad phonetic basis (Table 4.1), and the distributions of phone-level low-band energies in which phones are grouped on a fine phonetic basis. The p'th percentile φ^p and the standard deviation σ are calculated for each of the distributions. These values are shown in Tables 8.3 (duration) & 8.4 (energy).

The second inspection of the training data determines the distributions of the sum $z_{percentile}$ duration measures of all phones in the lhyme (Equation 4.7) of prominent and non-prominent syllables, for each unique number of phones that are found in a syllable lhyme. The calculation of the $z_{percentile}$ duration measures uses the statistics derived from the first inspection of the training data. The mean value is calculated for each of the distributions. These values are shown in the lower part of Table 8.3.

The integrated prosodic analysis system is used to assign sentential stress categories to the syllables in the test data (an open test). These categories are compared with those transcribed by hand (see Table 8.5). The transcriptions are in agreement for 79.1% of

Broad phonetic class	$\varphi_T^p(broad\_phone\_class)$ (ms)	$\sigma_T(broad\_phone\_class)$ (ms)
reduced monophthong	126.19	19.90
short monophthong	78.40	35.89
long monophthong	109.90	47.65
diphthong	126.79	48.58
sonorant	73.53	24.58
voiced obstruent	91.15	28.85
unvoiced obstruent	117.40	38.37
No. phones in lhyme, i	**$\overline{M_s}(i)$**	**$\overline{M_u}(i)$**
1	0.865	-1.483
2	0.321	-3.312
3	-1.043	-4.061
4	-3.172	-6.715

Table 8.3: Duration feature normalisation statistics

the syllables.

8.5.2 Bayesian classifier (AW-system)

Each syllable (derived using the acoustic-phonetic syllabification algorithm presented in Section 8.1.1) is characterised by four features:

- The integral of the low-band energy over the syllable nucleus (*ptpint*).
- The duration of the syllable nucleus (*sondur*).
- The maximum *FØ* in the entire syllable (*F0max*).
- The average Euclidean distance of frame-to-frame log-power spectra over the central half of the syllable nucleus (*spchave*).

A Bayesian classifier (Appendix B) is trained using the training data to model three categories of sentential stress — pitch accented "*PA*", stressed but unaccented "*s*", and unstressed "*u*". The Bayesian classifier is used to assign sentential stress categories to syllables in the test data (an open test). These categories are compared with those

Fine phonetic class	$\varphi_E^p(fine\_phone\_class)$ (dB)	$\sigma_E(fine\_phone\_class)$ (dB)
[ɪ]	-5.885	5.160
[i]	-7.768	3.898
[ʊ]	-7.195	3.374
[u]	-9.406	3.479
[ʊə]	-7.429	2.844
[ɪə]	-7.010	2.829
[eɪ]	-5.757	3.136
[əʊ]	-5.075	3.064
[ɔɪ]	-7.500	3.419
[ə]	-1.633	4.862
[ɔ]	-6.814	3.380
[aɪ]	-4.026	3.250
[ɛ]	-6.029	3.232
[ɛə]	-5.320	2.708
[aʊ]	-4.474	3.455
[ʌ]	-4.673	3.063
[ɜ]	-4.722	2.799
[ɒ]	-4.869	3.284
[a]	-5.171	2.644
[ɑ]	-4.788	3.908
[l]	-0.615	4.794
[r]	-0.390	5.738
[w]	-5.098	6.173
[j]	-6.702	3.780
[n]	-6.700	4.865
[m]	-5.801	4.450
[ŋ]	-8.446	4.290
[b]	-10.375	10.271
[ʒ]	-14.526	9.137
[v]	-9.645	6.205
[g]	-14.094	8.161
[ʤ]	-10.095	7.426
[ð]	-5.174	8.643
[d]	-11.093	8.686
[z]	-10.612	6.581
[s]	-19.290	4.813
[θ]	-12.988	6.254
[t]	-5.275	11.458
[ʧ]	-19.554	9.021
[f]	-16.002	4.676
[ʃ]	-19.851	4.494
[p]	-15.663	11.113
[k]	-10.607	10.747
[h]	-11.507	7.013

Table 8.4: Energy feature normalisation statistics

		PCB-system label			
		PA	*s*	*u*	total
Hand Label	*PA*	1264 (23.8%)	293 (5.5%)	280 (5.3%)	1837 (34.5%)
	s	78 (1.5%)	55 (1.0%)	128 (2.4%)	261 (4.9%)
	u	234 (4.4%)	100 (1.9%)	2887 (54.3%)	3221 (60.6%)
	total	1576 (29.6%)	448 (8.4%)	3295 (61.9%)	5319 (100.0%)

Correct classification rate = 4206/5319 (79.1%)

Table 8.5: PCB-system: Confusion matrix of prosodic transcription
(*PA* — pitch accented; *s* — stressed but unaccented; *u* — unstressed)

		AW-system label			
		PA	*s*	*u*	total
Hand Label	*PA*	938 (17.6%)	0 (0.0%)	899 (16.9%)	1837 (34.5%)
	s	99 (1.9%)	0 (0.0%)	162 (3.0%)	261 (4.9%)
	u	384 (7.2%)	0 (0.0%)	2837 (53.3%)	3221 (60.6%)
	total	1421 (26.7%)	0 (0.0%)	3898 (73.3%)	5319 (100.0%)

Correct classification rate = 3775/5319 (71.0%)
(Entropy score = 0.956)

Table 8.6: AW-system: Confusion matrix of prosodic transcription
(*PA* — pitch accented; *s* — stressed but unaccented; *u* — unstressed)

transcribed by hand (see Table 8.6). The transcriptions are in agreement for 71.0% of the syllables. No syllables are labelled as stressed but unaccented "*s*" by the Bayesian classifier, and syllables labelled as prominent ("*PA*" or "*s*") by hand are labelled as non-prominent ("*u*") more often than they are labelled as prominent by the Bayesian classifier.

8.5.3 Knowledge-based rules approach (JLH-system)

The duration and energy thresholds used in the knowledge-based rules approach proposed by Hieronymus (1989; 1991) must first be determined from the training data. To do this, the vowels are classified on a broad phonetic basis as either reduced [ə], short

[ɪ, ɨ, ɛ, a, ʌ, ʊ, ɒ, æ], long [i, ɜ, u, ɔ, ɑ, ʉ, ɚ] or a diphthong [eɪ, aɪ, ɔɪ, əʊ, aʊ, ɪə, ɛə, ʊə]. For each vowel-type, the distribution of durations is calculated with a fixed adjustment factor (0.6) being applied to all pre-pausal vowels. The percentage $p(vowel\_type)$ of vowels transcribed by hand as unstressed is determined for each vowel-type. The duration threshold for each vowel-type is given by the $p(vowel\_type)$'th percentile of the corresponding distribution. A distribution of maximum intra-vowel low-band energies is calculated across all vowels. The energy threshold, above which the intensity is regarded as contributing to the prominence of a vowel, is given by the p_{all}'th percentile of the all-vowel-types energy distribution, where p_{all} is the percentage of all vowels which are hand transcribed as unstressed.

Although the knowledge-based rules approach proposed by Hieronymus makes no distinction between reduced and short vowels in assigning a level of prominence, the reduced vowels are separated from the category of short vowels in determining the duration thresholds. This is done because all the reduced vowels in the training data are transcribed as unstressed and would offset the duration threshold if included.

The distributions of pre-pausal compensated durations and maximum intra-vowel low-band energies for the training data are shown in Figure 8.6. The duration and energy thresholds derived from these distributions are indicated by the dotted lines. The -20.0dB energy threshold is used by the knowledge-based rules approach to annotate all low intensity vowels as unstressed, regardless of other acoustic parameters.

The knowledge-based rules approach is used to assign sentential stress categories to the vowels in the test data (an open test). These categories are compared with those transcribed by hand (see Table 8.7). The transcriptions are in agreement for 72.6% of the syllables.

8.5.4 Comparison

It is important to determine if the differences in performance of these algorithms is statistically significant before any claims can be made about their relative efficacy. The McNemar test (Gillick & Cox, 1989) is used to decide whether or not the difference in the classification error rates between any two algorithms is statistically significant. The null hypothesis H_0 to test is that the true (but unknown) error rates of two algorithms

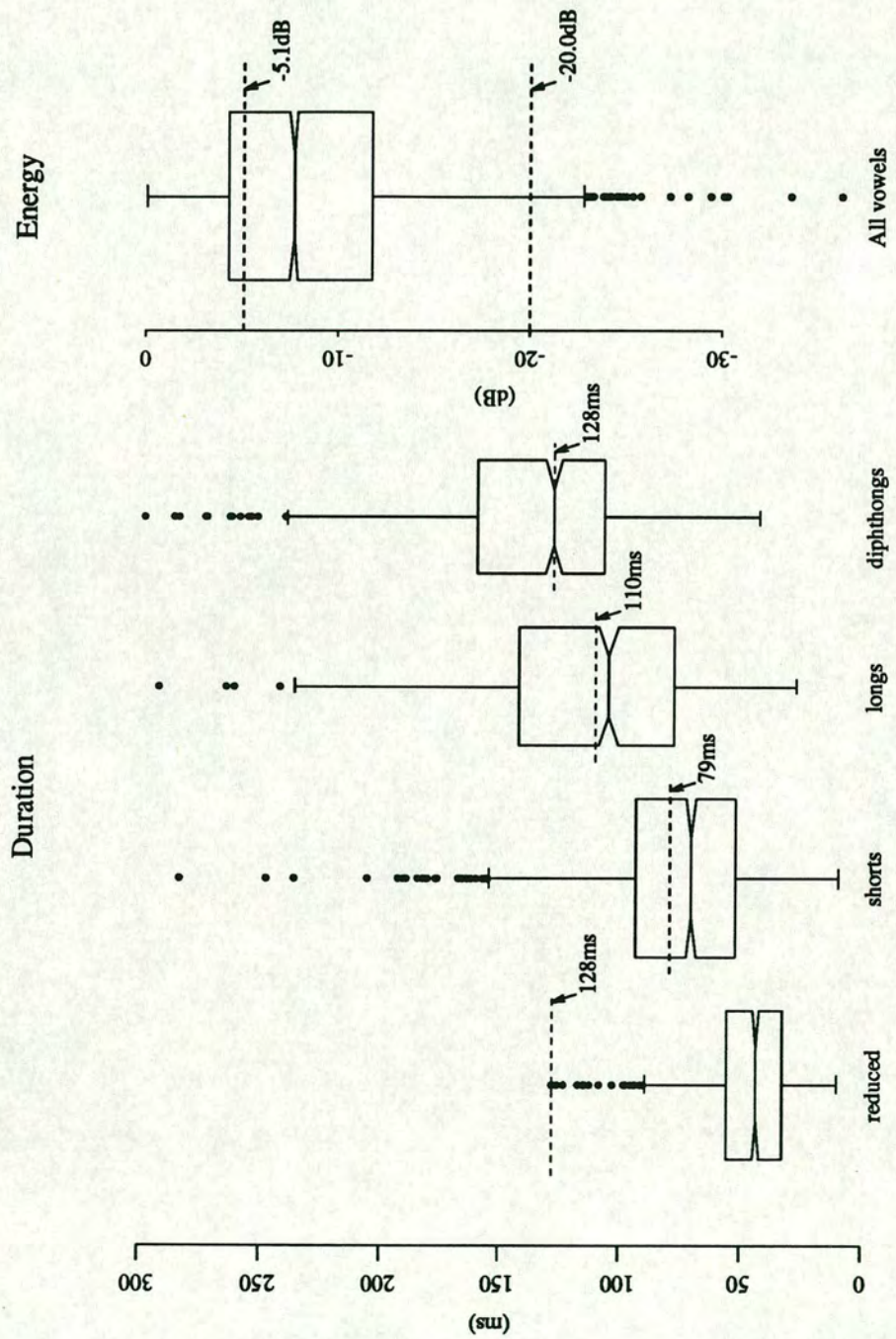


Figure 8.6: Parameter distributions for Hieronymus algorithm

		JLH-system label			
		PA	*s*	*u*	total
Hand Label	*PA*	1267 (23.8%)	328 (6.2%)	242 (4.5%)	1837 (34.5%)
	s	107 (2.0%)	58 (1.1%)	96 (1.8%)	261 (4.9%)
	u	461 (8.7%)	223 (4.2%)	2537 (47.7%)	3221 (60.6%)
	total	1835 (34.5%)	609 (11.4%)	2875 (54.1%)	5319 (100.0%)

Correct classification rate = 3862/5319 (72.6%)

Table 8.7: JLH-system: Confusion matrix of prosodic transcription
(*PA* — pitch accented; *s* — stressed but unaccented; *u* — unstressed)

are the same.

A binary decision is made as to whether an algorithm classifies a syllable correctly or incorrectly. A two-by-two cross-performance matrix is used to indicate the number of syllables which are correctly classified by two different algorithms, the number of syllables which are incorrectly classified by both algorithms, and the number of syllables which are correctly classified by one of the algorithms but incorrectly classified by the other algorithm (or visa versa). The cross-performance matrices for each two-way selection of the algorithms (the PCB-system, the AW-system and the JLH-system) are shown in Table 8.8.

The null hypothesis H_0 is tested by applying a two-tailed test to an observation drawn from a Binomial distribution. The probability P of the observation is calculated from a cross-performance matrix using the method described by Gillick & Cox (1989).

In comparing the error rates of the PCB-system and the AW-system, and the error rates of the PCB-system and the JLH-system, P is negligible. The null hypothesis is therefore rejected and there is evidence of a genuine difference. However, in comparing the error rates of the AW-system and the JLH-system, $P = 0.028$. The null hypothesis cannot by rejected at the 2% significance level.

The PCB-system and the JLH-system use the same pitch accent decision filter. There are fewer stressed but unaccented and unstressed syllables which are confused as being pitch accented by the PCB-system than by the JLH-system. (Compare the left-hand column of Tables 8.5 & 8.7.) This reduction in errors is due to the stylisation and

		AW-system	
		Correct	Incorrect
PCB-system	Correct	3307	899
	Incorrect	468	645

		JLH-system	
		Correct	Incorrect
PCB-system	Correct	3404	802
	Incorrect	458	655

		AW-system	
		Correct	Incorrect
JLH-system	Correct	3056	806
	Incorrect	719	738

Table 8.8: Cross-performance matrices of algorithm pairs

schematisation of the *FØ* contour in the PCB-system. The application of the extensive processing of the *FØ* contour appears to satisfy its goal of preventing microprosodic variations from being confused as pitch accents. There is, however, a slight increase in the number of pitch accented syllables which are confused as being unstressed by the PCB-system in comparison with the JLH-system. Some pitch accent related *FØ* variations may also be removed by the stylisation and schematisation processes.

There are fewer prominent syllables which are confused as being non-prominent syllables by the PCB-system than by the AW-system. There are also fewer non-prominent syllables which are confused as being prominent syllables by the PCB-system than by either the AW-system or the JLH-system. This marked reduction in errors is due to the normalisation of the acoustic parameters for non-prosodic aspects of speech in the PCB-system.

Methods of normalising acoustic parameters for non-prosodic aspects of speech, and the techniques of processing *FØ* contours to isolate the intonational component are proposed in this thesis. It is concluded from the above comparison that these techniques provide a statistically significant improvement in the performance of automatic prosodic analysis systems.

8.6 Conclusions

An automatic prosodic analysis system is described which integrates the unidimensional contributions made by features of duration, energy and fundamental frequency. The system produces an acoustic-phonetic representation of sentential stress patterns and intonation within a syllabic domain.

An algorithm is proposed to group phones automatically into syllables using acoustic-phonetic parameters. The syllabification algorithm is unique in its use of both a low-band energy contour and the segmentation (phone boundary and label information) of an utterance. Each syllable generated by this algorithm is associated with one local maximum in the low-band energy contour per group of potential syllable nuclei (vowels and syllabic consonants). It is assumed that such syllables can only ever be perceived as single prominent units in connected speech. This automatic syllabification from acoustic-phonetic parameters is shown to have a large level of agreement (95.2%) with a syllabification based on abstract phonological rules. The automatic syllabification algorithm has advantages over the syllabification based on phonological rules in that it is robust to errors in the segmentation of an utterance and that it groups phones according to the manner in which the speaker produces the utterance rather than according to some predicted manner. This is particularly important for the syllabification of an utterance produced by a non-native speaker, where the actual pronunciation may deviate from a pronunciation which may be predictable from phonological rules only for a native speaker.

Syllables are used in the extraction of duration and energy features and in the processing of fundamental frequency. The syllabic domain also ties together the information embedded within these three acoustic parameters.

The processing of fundamental frequency is based on the principle that an *FØ* contour has a composite structure (see Section 3.4). The piece-wise stylisation of an *FØ* contour aimed at eliminating microprosodic variations is illustrated. The piece-wise sections are abstracted to form *FØ* trajectory descriptors which are independent of speaker-dependent effects. This leads to a schematic representation of an *FØ* contour which provides an acoustic-phonetic representation of the intonation of an utterance. The analysis of an *FØ* schema by a pitch accent decision filter provides judgements of syllable accentuation. The processing of raw *FØ* contours to form *FØ* schemata reduces the

number of unaccented syllables which are erroneously classified as being pitch accented by the pitch accent decision filter.

The unidimensional contributions to the perceived prominence of a syllable made by features of duration and energy are captured by a peak-picking algorithm. The information captured by the peak-picking algorithm and the syllable accentuation information are combined to produce a description of the sentential stress patterns of an utterance.

The intergrated prosodic analysis system yields a 79.1% agreement level with sentential stress categories transcribed by hand. This is shown to be a statistically significant improvement over the agreement levels yielded by two former algorithms. This level of agreement is comparable with the 83.1% agreement between transcriptions of syllable prominences made independently by two phoneticians in the Lancaster/IBM spoken English corpus (Pickering *et al.*, 1994).

Chapter 9

Application to Computer Aided Pronunciation Teaching

This Chapter describes how the integrated prosodic analysis system outlined in Section 8.4 can be applied to computer aided pronunciation teaching. Section 9.1 gives a general description of the framework of the SPELL[1] system (Lefèvre *et al.*, 1992; Hiller *et al.*, 1993) within which the proposed automatic prosodic analysis system can operate. An example of the operation of the automatic prosodic analysis within the framework of SPELL is described in Section 9.2.

A short discussion of possible extensions to the work presented in this thesis, is presented in Section 9.3.

9.1 Framework of the SPELL system

SPELL is a feasibility study to develop a tool for teaching prosodic and segmental aspects of pronunciation to non-native students of English, French and Italian. The SPELL workstation is an autonomous teaching system which is used by a student without the need for a language teacher to organise the pronunciation tasks. The system is aimed for use by students who do not need to be taught the basics of a foreign language as well as pronunciation. A student is provided with both audio and visual aids. The audio aids enable a student to listen to the pronunciation of utterances. The utterances presented to a student may be either the ideal target pronunciation spoken by a native

[1]SPELL (Interactive System for Spoken European Language Training) is a project supported by the European Community's ESPRIT programme, under contracts No.5192 & No.7153.

talker, or a resynthesised version of the student's voice with emphasis placed on a specific feature of its pronunciation. The visual aids enable a corrective diagnosis of a student's pronunciation to be presented to the student without the student requiring a knowledge of the underlying phonological theory or phonetic concepts on which the diagnosis is based. The objective of the SPELL system is to improve a student's intelligibility rather than to promote fluency in a foreign language or to encourage near-perfect mimicry of a native speaker.

Foreign language pronunciation is taught in the SPELL system under a teaching methodology called the DELTA paradigm. This paradigm constitutes four steps:

- Demonstate — A student is presented with audible utterances spoken by a native talker which highlight a particular feature of pronunciation.
- Evaluate Listening — The student performs a number of listening tests to ensure that the pronunciation feature which was demonstrated in the previous step can be perceived by the student. If the student fails to exhibit competence, the demonstration phase is repeated.
- Teach — A feature of pronunciation is taught to the student. This involves presenting an exemplar utterance to the student and requesting the student to reproduce the utterance whilst concentrating on the pronunciation feature of interest. Quantitative feedback and a corrective diagnosis are provided in order to modify inadequacies in the student's reproduction of the utterance.
- Assess — The student's ability to generate a particular feature of pronunciation is evaluated. This may involve requesting the student to pronounce a number of previously unheard utterances containing the pronunciation feature of interest. If the student completely fails to exhibit competence, the demonstration phase is returned to. If the student shows only some ability, the teaching phase is repeated.

In teaching prosodic aspects of a language with this methodology, an automatic prosodic analysis system is required within the teaching and assessment phases of the DELTA paradigm.

There is an enormous degree of variation in the trajectories of $F\emptyset$ contours. It is therefore not possible to teach intonation simply by encouraging a student to reproduce

exact imitations of isolated contours. A student needs to be familiarised with patterns or models of *FØ* trajectories which can be generalised to utterances of the same type, and needs to acquire an ability to select a particular pattern or model to convey a specific linguistic function (Rooney *et al.*, 1992). Therefore, in the teaching of intonation within the SPELL framework, prompts are used to provide a communicative context for the student and to elicit an appropriate pattern or model of *FØ* trajectories in the student's response. In addition, the segmental structure of an utterance is associated with the *FØ* trajectories which make up an intonation pattern. This is required to ensure that pitch accents are realised at clearly defined locations in an utterance.

The intonation teaching within the SPELL framework is primarily limited to two patterns of *FØ* trajectories; one pattern is typically used in syntacticly simple (subject-verb-object or subject-verb-adverb) declarative statements and *wh*-questions, and the other pattern is used in polar (*yes/no*) questions.

9.2 Example of automatic prosodic analysis applied to the SPELL framework

In this example, an intonation pattern commonly applicable to syntacticly simple declarative statements in English is being taught to a student whose mother-tongue is Italian. One of the chief characteristics of the intonation pattern is a falling pitch that is associated with the focus of the statement.

The exemplar utterance is, "I enjoy swimming these days." This statement is prompted with the question, "What sports do you do?" in order to provide a plausible context for the student and to elicit an emphasis on the word *swimming*. In contrast, the prompt question, "Do you hate swimming these days?" may be used to elicit an emphasis on the word *enjoy*.

In the teaching phase of the DELTA paradigm, the student is provided with a visual orthographic transcript of the prompt question. This is accompanied by a recording of the question which is played to the student. An audible version of the declarative statement pronounced by a native English speaker (referred to as the *teacher*) is then presented. The student may repetitively listen to the teacher's pronunciation of the state-

ment. An automatic prosodic analysis of the teacher's utterance is also presented to the student as a visual aid. The student attempts to reproduce the statement with the same sentential stress pattern and intonation as exhibited in the ideal target pronunciation demonstrated by the teacher. An automatic prosodic analysis of the student's utterance is performed, by the system outlined in Section 8.4, using the same acoustic parameter normalisation statistics that are used in the analysis of the teacher's utterance. This ensures that a syllable pronounced by the student is automatically transcribed as prominent only if the relative variations in the duration and energy measures are analogous to those in the speech of a native English talker. The automatic prosodic analysis of the student's utterance is presented such that the beginning and end points are aligned with the analysis of the teacher's utterance (see Figure 9.1).

In terms of the configuration theory, the target pronunciation of, "I enjoy swimming these days," includes a falling nuclear pitch movement beginning on the first syllable of the word *swimming*. In the pronunciation demonstrated by the teacher, the word *I* is stressed but unaccented (part of the prehead) and the second syllable of *enjoy* is accented (the head). The word *days* is also stressed but unaccented (part of the tail).

Although the *FØ* trajectory within the student's pronunciation of *I* is falling (in contrast to the rising trajectory in the target pronunciation) the syllable is categorised as stressed but accented. The pronunciation of this word is therefore satisfactory and the difference in the shape of the trajectories is not brought to the attention of the student during the diagnosis of the pronunciation. There is, however, an underlying error in the phonetic realisation of the diphthong /aɪ/. The automatic segmentation of the Italian student's pronunciation of *I* is represented as two monophthongs [a i]. These segmental aspects of pronunciation are taught as a separate issue within the SPELL framework.

Two errors occur in the prosodic aspects of the student's pronunciation. The student successfully places the nuclear pitch movement on the first syllable of the word *swimming*. This is automatically detected as being an accented syllable, however a rising *FØ* trajectory is associated with its realisation. The student also accentuates *days* with a continuation rise. In addition, the word *days* is associated with a segmental pronunciation error in the diphthong /eɪ/. The automatic segmentation of the Italian student's pronunciation of *days* is represented as [d ɛ i z]. The student's attention can be drawn

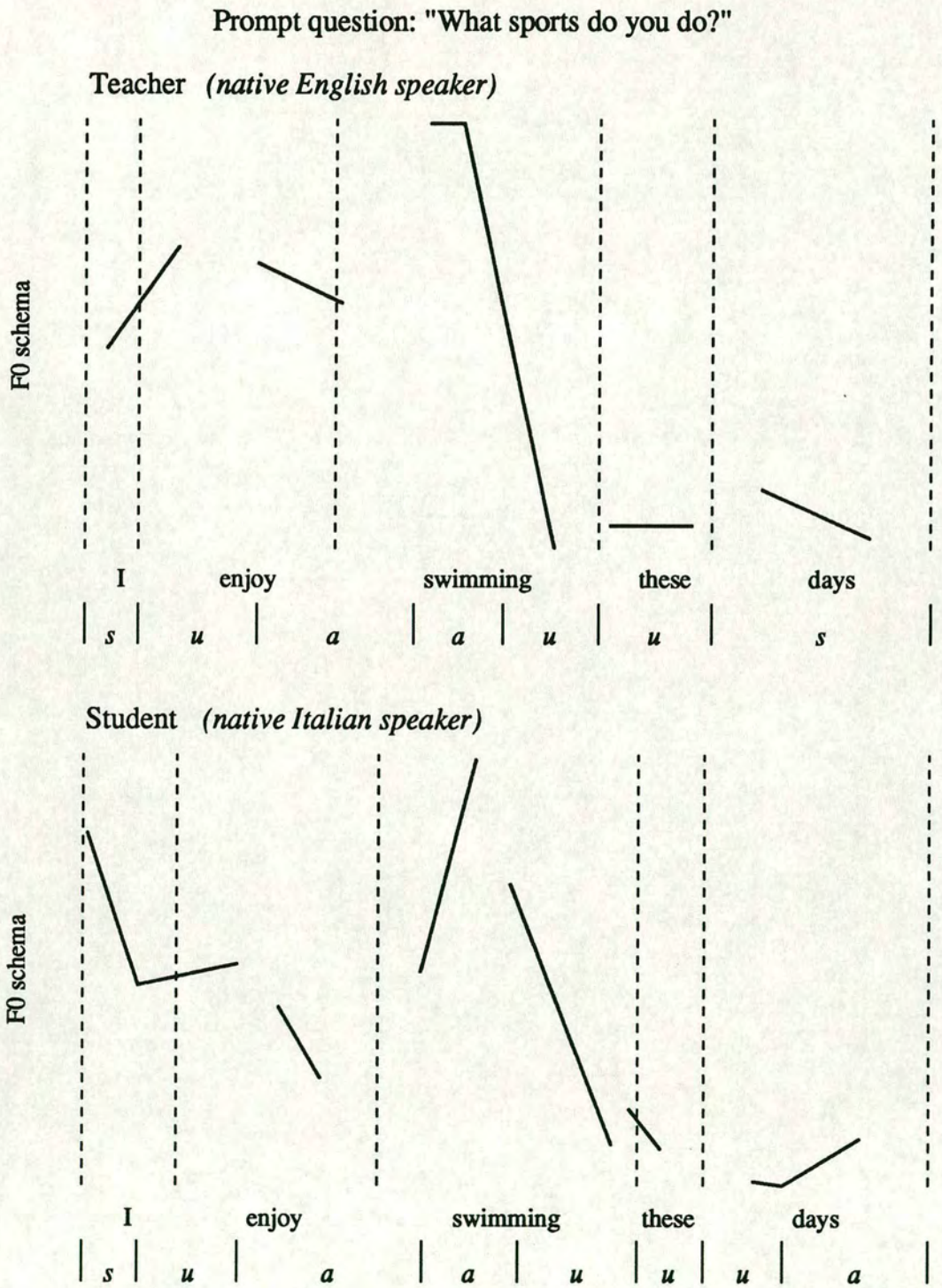


Figure 9.1: Example of prosodic analysis in pronunciation teaching

to the prosodic pronunciation errors by highlighting differences in the *FØ* schemata and the transcription of sentential stress.

The analyses of duration and energy measures are embedded in the sentential stress transcription. If a syllable is detected as being unstressed when the target pronunciation is to produce a prominent syllable, then a student may be instructed to make a specific syllable "longer" and/or "louder" in order to achieve the required prominence.

The acoustic-phonetic representation makes no references to any phonological theories. It is therefore suitable for language students who do not need a proficient knowledge of linguistic theory.

9.3 Further work

There are areas of the integrated prosodic analysis system where the underlying research may be extended. The aim of any further work would be to ensure that a foreign language learner is advised to alter the pronunciation of an utterance only in parts where it can convey a semantically contrastive function.

In the low-level analysis of the acoustic parameters, additional normalisation is required to compensate for contextual influences on the duration and energy of phones. A greater understanding of the effects of contextual assimilation on vowels may enable a vowel quality measure to be devised which is correlated with the perceived prominence of a syllable.

The unidimensional contributions of the acoustic parameters are integrated after the normalisation of non-prosodic aspects of speech. Integration of the acoustic parameters at a lower level of the analysis may enable, for example, duration measures to assist in the normalisation of energy measures or visa versa. A reduction in energy at the end of an intonational phrase may assist in normalising a duration measure for pre-pausal lengthening effects.

The abstraction of an *FØ* schema to a phonological representation (in terms of high and low tones) would enable the intonation of a student's speech to vary only in places where it does not convey a semantically contrastive function. Such abstraction must be reliable and the phonological representation must have a firm theoretical foundation. Moreover, the phonological representation must consider the mother-tongue of the

foreign language learner.

Other aspects of prosodic structure need to be automatically transcribed in order to teach a foreign language learner rhythm and syllable timing differences between languages. The automatic transcription of break indices (which represent the degree of prosodic coupling between neighbouring words) would assist in the analysis of rhythm.

Finally, the automatic prosodic analysis of the speech needs to be performed as quickly as possible to prevent a user of a pronunciation teaching system from growing impatient. In the integrated prosodic analysis system presented here, there is a bottleneck in the analysis of an $F\emptyset$ contour. The least median of squares regression is time consuming even though a Monte Carlo type speed-up technique is applied. Methods of reducing the computational work-load involved in the analysis may be worthy of investigation.

Chapter 10

Summary and Conclusions

This thesis presents research towards the automatic analysis of prosodic events in utterances of English spoken by native and non-native talkers. The research concentrates on the automatic analysis of speech to generate an acoustic-phonetic representation of sentential stress and intonation. The acoustic-phonetic representation of prosody is demonstrated as a tool for computer aided pronunciation teaching.

It is argued in Chapter 2 that prosodic aspects of speech need to be explicitly taught to students who wish to communicate competently in a foreign language. In summary, the argument is that prosodic aspects of speech need to be taught to a student because prosody has functional differences between languages, and the acoustic-phonetic realisation of prosodic aspects of speech also differs between languages. Explicit teaching is required because a student's comprehension and intelligibility can be influenced by prosodic features. Inter-language differences are illustrated in exemplar languages.

In teaching the prosody of a foreign language, it is suggested that a formalised phonological theory of prosody should not be explicitly taught to a foreign language learner. This suggestion is made by taking the analogy that a formalised grammar is usually not explicitly taught to a student who needs to acquire a knowledge of the grammatical structures of a foreign language. However, phonological theories of prosody are needed to design courseware by which a student can learn stress patterns and intonation.

Ideally, the prosodic composition of the speech of a foreign language learner should be described in a manner which allows any linguistically significant deviations from a near-native pronunciation to be determined. The theories of intonational phonology aim to describe a minimal set of pitch variations which are semantically contrastive. A

comparison of the prosodic composition of a student's speech and a native's speech can ideally be made by describing the prosody with phonological models. Such a comparison would allow a student's pitch to differ from that of a native speaker only at times where it does not carry any semantically contrastive function. A corrective diagnosis could then be offered to a student in order to aid the student's pronunciation.

However, it is argued in Chapter 2 that phonological representations of prosodic structure are language-specific. The prosodic composition of a student's speech can be highly influenced by mother-tongue interference and by a student's stereotypic images of prosody in the target language, thereby rendering a language-specific phonological representation of prosody inappropriate. Moreover, languages vary in the way acoustic characteristics of speech are modified to manifest prosodic aspects of speech. It is therefore proposed in Chapter 2, that the only secure means to describe prosody for foreign language teaching lies in an acoustic-phonetic representation. It is assumed that a higher level of analysis can be used to determine which of the detectable phonetic differences do have and do not have a semantically contrastive function in a specific language.

In order to derive an acoustic-phonetic representation of prosody from a speech signal, it is necessary to identify acoustic parameters which are correlated with prosodic aspects of speech. In Chapter 3, a review of research conducted to identify acoustic parameters indicates that duration, energy, fundamental frequency and vowel quality are the key correlates of prosodic features. There is relatively little consideration made in the literature to the effects of vowel quality on the perception of (lexical or sentential) stress, compared with the extensive research relating duration, energy and *FØ* with stress.

Studies of the acoustic correlates of stress propose a variety of different features for duration and energy. Duration features include the duration of the vocalic portion of a syllable and the duration of an entire syllable. Energy features include the peak amplitude in a syllable, the amount of fall from a peak in the amplitude envelope of a syllable, the integral of the peak-to-peak amplitude over the vocalic portion of a syllable, the average low-band syllable energy, and the maximum low-band vowel energy.

Chapters 4, 5, 6 & 7 concentrate on optimising the extraction of acoustic correlates required for the prosodic analysis of English. The duration and energy measures of phonetic units are dependent upon the definition of the phonetic units. A number of

different phonetic units related to syllable structure are investigated. The phonetic units investigated are the syllable nucleus, the syllable rhyme, the syllable lhyme, the entire syllable and a nucleus-to-nucleus unit. Duration and energy measures are made from these units. It is asserted that the duration and energy measures are influenced both by prosodic and by non-prosodic aspects of speech. Normalisation techniques are applied to the measures in order to compensate for a number of non-prosodic factors, including variations in the measures which are due to phone-type, syllable structure, syllable length in terms of number of phones, and the prominence level of syllables within the training data. A novel $z_{percentile}$-transform is proposed for the normalisations.

There are 213 different duration features investigated in Chapter 4. It is shown that factors other than sentential stress cause the duration of phones to the right of a syllable nucleus (but within the same syllable) to vary more than the duration of phones to the left of the nucleus. The syllable lhyme is defined as the unit which constitutes the onset and the nucleus of a syllable; hence the syllable lhyme contains the parts of a syllable which are least affected by factors other than sentential stress. The optimal duration feature, of those investigated, is related to the syllable lhyme with sum $z_{percentile}$ duration measures trained on broad phonetic classes without smoothing of the phone-level duration contour and with normalisation of the number of phones in the syllable lhyme. The optimal energy feature, out of 113 energy features investigated, is related to the syllable nucleus with $z_{percentile}$ low-band energy measures trained on fine phonetic classes.

A simple peak-picking algorithm is used to determine the contribution made by each acoustic parameter to the degree of syllable prominence. The optimal duration feature is used to label syllables as either prominent or non-prominent and yields a 77.8% agreement level with prominence labels based on human perception (in an open test). The optimal energy feature is used to label syllables as either prominent or non-prominent and yields an 84.0% agreement level with the prominence labels based on human perception.

A number of measures of vowel quality are proposed in Chapter 6 which are based on the assumptions that prominent syllables are well articulated and are less affected by contextual assimilation than non-prominent syllables, and that these properties are

reflected in the nuclei of syllables. A Bayesian classifier is used to obtain a measure of how close a vowel is to its respective vowel target model. This distance measure is used to label syllables spoken by a native English speaker as either prominent or non-prominent and yields a 71.6% agreement level with prominence labels based on human perception. It is argued that the distance measure cannot be used as a correlate of stress in utterances spoken by non-native speakers of English because of vowel pronunciation errors. A number of alternative measures are investigated as features to characterise the degree of vowel stability. The average Euclidean distance of frame-to-frame log-power spectra is used as a measure of the degree of spectral change in a vowel due to contextual assimilation. This vowel stability measure is used to label syllables as either prominent or non-prominent and yields a 68.8% agreement level with prominence labels based on human perception. A higher level of agreement is obtained by using a $z_{percentile}$-transformed, normally-distributed random variable as an input parameter to a peak-picking algorithm. The application of the vowel stability measure as a correlate of stress is therefore inappropriate. The use of phonological rules to transcribe syllables as non-prominent under certain conditions, is also dismissed.

Fundamental frequency is identified as an acoustic correlate of stress, and is used in algorithms (reviewed in Section 3.5) to automatically transcribe the location of prominent syllables in isolated words and in connected speech. In addition, fundamental frequency is principally related to the intonation of an utterance. Nevertheless, many of the reviewed algorithms use a measure of *FØ* as a key acoustic correlate of stress with little or no consideration of the intonational role of *FØ*. However, *FØ* can be used as a secondary cue to the location of prominent syllables in connected speech because pitch accents are observed to fall on prominent syllables.

The fundamental frequency of a speech waveform must be determined as an initial process for prosodic analysis, since it is associated with both sentential stress and intonation. The performance of *FØ* determination algorithms (a selection of which are reviewed in Section 3.2) must be considered with respect to their application in systems of prosodic analysis, and errors arising from the malfunctions of FDAs must be prevented from propagating into the subsequent prosodic analysis of speech. A number of modifications to the super resolution *FØ* determination algorithm are proposed in

Section 7.1. The modifications reduce the occurrence of errors involved in the extraction of *FØ* such that it is optimised for prosodic analysis, relative to the selection of other algorithms designed to determine *FØ*. A novel de-step filter is proposed to post-process an *FØ* contour generated by an FDA in a way which further reduces the occurrence of discontinuity errors commonly observed in *FØ* contours.

The processing of fundamental frequency is based on the principle that an *FØ* contour has a composite structure (Section 3.4). In particular, the *FØ* contour of an utterance is affected by the talker's anatomy and physiology, the segmental content of an utterance, cycle-to-cycle jitter (*FØ* perturbation) and errors involved in its determination from the speech waveform. The piece-wise stylisation of an *FØ* contour aims to eliminate microprosodic variations. An algorithm to stylise an *FØ* contour into linear piece-wise sections with respect to the syllables of an utterance, is illustrated in Section 8.2.1. The piece-wise sections are abstracted to form *FØ* trajectory descriptors which are independent of speaker-dependent effects. This leads to a schematic representation of an *FØ* contour which provides an acoustic-phonetic representation of the intonation of an utterance. The *FØ* trajectories of a schema are associated with syllables and a pitch accent decision filter analyses the trajectories to provide judgements of syllable accentuation. The integration of syllable information with the stylisation and schematisation of an *FØ* contour is an innovative approach. The processing of raw *FØ* contours to form *FØ* schemata reduces the number of unaccented syllables which are erroneously classified as being pitch accented by the pitch accent decision filter. The judgements of syllable accentuation provided by the pitch accented decision filter yields a 72.7% agreement level with syllables labelled as pitch accented by hand.

Syllables are used in the extraction of duration and energy features and in the processing of *FØ*. The syllabic domain also ties together the information embedded within these three acoustic parameters. Furthermore, prosodic aspects of speech are described in a syllabic domain. The automatic syllabification of speech is therefore an important part of prosodic analysis. Algorithms reviewed in Section 3.3 are devised to partition a speech signal into syllable-sized units by using measures of sonority based on low-band energy and by using spectral characteristics to locate boundaries between adjacent sonorants. The reviewed syllabification algorithms are applied to speech recognition sys-

tems. Reliable information about the segmental content of an utterance is therefore not available to them. In the application of prosodic analysis for computer aided pronunciation teaching, however, reliable segmental information is available and may be used to enhance the syllabification process.

An algorithm is proposed in Section 8.1.1 to group phones automatically into syllables using acoustic-phonetic parameters. The syllabification algorithm is unique in its use of both a low-band energy contour and the segmentation (phone boundary and label information) of an utterance. Each syllable generated by this algorithm is associated with one local maximum in the low-band energy contour per group of potential syllable nuclei (vowels and syllabic consonants). It is assumed that such syllables can only ever be perceived as single prominent units in connected speech. This automatic syllabification from acoustic-phonetic parameters is shown to have a large level of agreement (95.2%) with a syllabification based on abstract phonological rules. The automatic syllabification algorithm has advantages over the syllabification based on phonological rules in that it is robust to errors in the segmentation of an utterance and that it groups phones according to the manner in which the speaker produces the utterance rather than according to some predicted manner. This is particularly important for the syllabification of an utterance produced by a non-native speaker, where the actual pronunciation may deviate from a pronunciation which may be predictable from phonological rules only for a native speaker.

The unidimensional contributions to the perceived prominence of a syllable made by features of duration and energy are captured by a peak-picking algorithm. The pitch accent decision filter provides judgements of syllable accentuation. The information captured by the peak-picking algorithm and the syllable accentuation information are combined to produce a description of the sentential stress patterns of an utterance, in a syllabic domain. This yields a 79.1% agreement level with sentential stress categories transcribed by hand. This is shown to be a statistically significant improvement over the agreement levels yielded by two former algorithms.

The former algorithms for prosodic analysis reviewed in Chapter 3 have not addressed the problems of determining the fundamental frequency of speech and have not comprehensively addressed the need to normalise for non-prosodic variations in the acoustic

parameters which are used in the prosodic analysis. The system of prosodic analysis described in this thesis addresses the problems related to the determination of the fundamental frequency of speech and focuses on techniques of normalising for variations in acoustic parameters which are due to non-prosodic aspects of speech.

The integrated prosodic analysis system proposed in this thesis is shown, in Chapter 9, to produce prosodic descriptions which are useful in comparing the prosodic aspects of the speech of a non-native learner of English with the speech of a native English talker.

Appendix A

Least Median of Squares Regression

A highly robust method of fitting a linear regression model to a set of observations (including spurious samples) is least median of squared residuals (LMedS) regression analysis; introduced by Rousseeuw (1984).

Consider a set of N observations, $\{i, y_i\}_{i=1}^{N}$.

For all pairs of observations, $\{P_u, P_v \mid u \in 1,\ldots,N; v \in 1,\ldots,N; P_u = (u, y_u)\}$ a linear model is computed. There are $K = {}_N\mathbf{C}_2 = \frac{N!}{2!(N-2)!}$ such pairs of observations. The linear model calculated from the k'th pair of observations is represented as,

$$z'_k(i) = a_k.i + b'_k \text{ where } \begin{cases} a_k = \dfrac{y_u - y_v}{u - v} \\ b'_k \text{ need not be calculated} \end{cases} \tag{A.1}$$

The point of intersection, b'_k, is disregarded since it is not a robust value. Let a simplified version of the model $z_k(i) = a_k.i$. The set of residuals (offset by b'_k) $\{\alpha_i = y_i - z_k(i)\}_{i=1}^{N}$ is computed and a mode-seeking algorithm is applied to it.

One method of estimating the mode (the sample value which occurs most frequently) of a distribution is named "estimating the rate of an inhomogeneous Poisson process by $(N/2)$'th waiting times," (Press *et al.*, 1988, chapter 13.3). The set $\{\alpha_i\}_{i=1}^{N}$ is initially sorted into ascending order to give $\{\hat{\alpha}_i\}_{i=1}^{N}$. The smallest window is then sought which accommodates half of the data ($N/2$ points) in the distribution of sorted offset residuals.

That is to say, δ_k is determine as,

$$\delta_k = \frac{1}{2} \cdot \min_{i=1 \rightarrow \lfloor N/2 \rfloor} (\hat{\alpha}_{i+\lceil N/2 \rceil} - \hat{\alpha}_i) \tag{A.2}$$

50% of the observations exist within the window and 50% exist outside the window. The evaluation of Equation A.2 is equivalent to shifting this window through the set of sorted offset residuals, $\{\hat{\alpha}_i\}_{i=1}^{N}$, to locate the area of highest density within its distribution. Let this point of highest density be located when $i = j$. A robust value of the point of intersection, b_k, is given by the mid-point of the window when $i = j$,

$$b_k = \frac{\hat{\alpha}_{j+\lfloor N/2 \rfloor} + \hat{\alpha}_j}{2} \tag{A.3}$$

The magnitude of the residuals within the window is always less than δ_k. Let $\hat{\alpha}_{[i]}$ denote the 50% of sorted offset residuals which reside within the window. Thus, $(b_k - \hat{\alpha}_{[i]})^2 < \delta_k^2$. Since $\hat{\alpha}_{[i]}$ corresponds to 50% of the data, then δ_k^2 is the median[1] of the squared residuals by definition. Hence $\delta_k = \sqrt{\text{med}(b_k - \hat{\alpha}_i)^2}$.

This process of determining a_k (from Equation A.1), b_k (from Equation A.3) and δ_k (from Equation A.2) is repeated for all k, giving $\{a_k, b_k, \delta_k\}_{k=1}^{K}$. It is then determined which pair of observations gives the minimum median squared residuals.

$$k_{min} = \underset{k=1 \rightarrow K}{\text{argmin}} (\delta_k^2) \tag{A.4}$$

If significant Gaussian noise is present simultaneously with numerous outliers, then the LMedS estimates, $\{a_{k_{min}}, b_{k_{min}}\}$ become less reliable. Thus, the LMedS algorithm is used for outlier detection and least (mean of) squares regression analysis is employed.

The data, modelled by $z''_{k_{min}}(i) = a_{k_{min}} . i + b_{k_{min}}$, has a standard deviation, $\hat{\sigma}$, from the model given by (Rousseeuw & Leroy, 1987),

$$\hat{\sigma} = 1.4826 \left(1 + \frac{5}{N-2}\right) . \delta_{k_{min}} \tag{A.5}$$

[1]The *median*, φ_{50}, of a set of samples is a number such that at least 50% of the sample values are smaller than or equal to φ_{50} and also at least 50% of those values are larger than or equal to φ_{50}. If there is more than one such number (in which case there will be an interval of them), the median is defined as the average of the numbers (midpoint of the interval).

All points for which the magnitude of the residual, $|y_i - (a_{k_{min}}.i + b_{k_{min}})|$, is less than or equal to $2.5\hat{\sigma}$ are taken to be *inliers* and all others samples are taken to be *outliers*. A linear model is then determined for all the inliers using least (mean of) squares regression analysis (see Press *et al.*, 1988, chapter 4).

The median computations are applied over the whole data set. Therefore, the LMedS estimator remains reliable if less then half of the observations are contaminated by outliers; ie. it has a 0.5 asymptotic breakdown point. The time-complexity of the algorithm, however, is high. There are $O(n^2)$ pairs and for each of them a sorting $O(n \log_2 n)$ is required. Thus, the LMedS algorithm has a time-complexity of the order $O(n^3 \log_2 n)$.

It is possible to reduce the time-complexity of the algorithm by applying a *Monte Carlo* type speed-up technique. When ${}_N\mathbf{C}_2$ is very large, a random subset of pairs of observations (m pairs) are taken, reducing the time-complexity to $O(mn \log_2 n)$. In order for the LMedS estimates to be reliable, it is necessary for at least one candidate model, z_k to carry the correct parameter values. If this condition is not met, the LMedS estimator loses its robust properties. Note, however, that in practice, more than one such candidate model is preferred, because noise will be present amongst the inliers.

Let ϵ be the fraction of data contaminated by outliers, and m be the number of randomly selected pairs. The probability of all m pairs containing at least one outlier is $P = [1 - (1 - \epsilon)^2]^m$ and thus $Q = 1 - P$ is the probability that at least one of the m pairs contains no outliers. For $\epsilon < 0.5$, taking $m = 3000$ pairs (${}_N\mathbf{C}_2 > 3000$ for $N > 77$), gives $P \approx 0$, $Q \approx 1$. Hence, there is a high probability of maintaining a high breakdown point when the Monte Carlo type speed-up technique is applied (Rousseeuw & Leroy, 1987).

Appendix B

Bayesian Classification

Consider a number of categories, $\{C_i \mid i \in 1,\ldots,I\}$, for a population of objects where each object is characterised by a set of p features. The aim of a classifier is to partition the p-dimensional feature space into regions, one region for each category, with the minimum probability of error.

A classification rule can be regarded as being optimal if the proportion of objects that are misclassified to any of the possible categories other than that which it belongs, is minimised by its use. Bayes' rule aims to minimise this classification error by assigning an object to a category with the highest *conditional probability*.

Bayes' rule (see, for example, Duda & Hart, 1973; James, 1985): An object is assigned to category C_i if,

$$P(C_i|\boldsymbol{x}) > P(C_j|\boldsymbol{x}) \qquad \forall\, j \neq i \tag{B.1}$$

where $\boldsymbol{x}$ is a vector of p feature measurements which characterise the object.

If the conditional probability $P(C_i|\boldsymbol{x})$ is a maximum for more than one category, the object is randomly assigned to one of the categories for which $P(C_i|\boldsymbol{x})$ is a maximum.

In practice, it is possible to estimate the probability $P(\boldsymbol{x}|C_i)$ of getting a particular set of features $\boldsymbol{x}$ given a sample of data from a particular category C_i. Bayes' theorem (see, for example, James, 1985) relates the *a posteriori* probability $P(C_i|\boldsymbol{x})$ to the probability $P(\boldsymbol{x}|C_i)$ and the *a priori* probability $P(C_i)$ of category C_i existing in the population:

$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i).P(C_i)}{\sum_{\text{all } i} P(\boldsymbol{x}|C_i).P(C_i)} \tag{B.2}$$

Bayes' rule can be rewritten by substituting Equation B.2 into Equation B.1 as:

Assign an object to category C_i if,

$$P(\boldsymbol{x}|C_i).P(C_i) \; > \; P(\boldsymbol{x}|C_j).P(C_j) \qquad \forall\, j \neq i \tag{B.3}$$

This rule is optimal for classification but requires a large training sample set to determine $P(\boldsymbol{x}|C_i)$ $\forall i$ and the *a priori* probabilities $P(C_i)$ $\forall i$ (which may be estimated by the proportion of each category in the sample set).

Let us assume that the measurements being used for classification form a multivariate Normal distribution, thus,

$$P(\boldsymbol{x}|C_i) \; = \; \frac{1}{(2\pi)^{p/2}.|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] \tag{B.4}$$

where,

$$\begin{aligned} \boldsymbol{\Sigma}_i &= E\left[(\boldsymbol{x}-\boldsymbol{\mu}_i).(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\right] \text{ (covariance matrix)} \\ \boldsymbol{\mu}_i &= E\left[\boldsymbol{x}_i\right] \text{ (mean vector)} \end{aligned}$$

Substituting Equation B.4 into Equation B.3, taking the natural logarithm, cancelling common terms, multiplying by -2 and reversing the inequality, gives Bayes' rule for multivariate Normal data as: Assign an object to category C_i if,

$$Q_i(\boldsymbol{x}) \; < \; Q_j(\boldsymbol{x}) \qquad \forall\, j \neq i \tag{B.5}$$

where the *quadratic discriminant score*, $Q_i(\boldsymbol{x})$, is given by,

$$Q_i(\boldsymbol{x}) \; = \; \ln|\boldsymbol{\Sigma}_i| + (\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - 2\ln(P(C_i)) \tag{B.6}$$

The above rule enables an object to be classified to a category on the basis of a vector of features, $\boldsymbol{x}$, given the mean values of the features, a covariance matrix of the features for each of the possible categories, and the *a priori* probability of an object existing in a given category (all of which can be determined from a training set). If the values of the features form a Normal distribution, the resultant classification will have a minimum probability of error.

Note that,

$$\exp\left[-\frac{Q_i(\boldsymbol{x})}{2}\right] \propto P(\boldsymbol{x}|C_i).P(C_i) \tag{B.7}$$

Thus, if by applying Bayes' rule expressed in Equation B.5, an object is assigned to category C_i, then the *a posteriori* probability that the object has been assigned to the correct category (substituting Equation B.7 into Equation B.2),

$$P(C_i|\boldsymbol{x}) = \frac{\exp\left[-\frac{Q_i(\boldsymbol{x})}{2}\right]}{\sum_{\text{all } j} \exp\left[-\frac{Q_j(\boldsymbol{x})}{2}\right]} \tag{B.8}$$

The measure adopted to evaluate the performance of the classification is the entropy (Shannon & Weaver, 1949; Pierce, 1962) associated with the task of correctly identifying the true category of the input object[1]. An entropy score can be determined from the *a priori* and the *a posteriori* probabilities of each class to give an indication of the effectiveness of the classification. The entropy score, H, is given as,

$$H = \sum_{\text{all } i}\left[P(C_i).\frac{1}{N_i}\sum_{n=1}^{N_i} -\log_2 P(C_i|\boldsymbol{x}_{i,n})\right] \tag{B.9}$$

where N_i represents the number of objects of category C_i in the sample set and $\boldsymbol{x}_{i,n}$ is the n'th feature vector characterising an object in the sample set which is known to be of category C_i. The *a posteriori* probabilities, $P(C_i|\boldsymbol{x}_{i,n})$, are calculated from Equation B.8.

If the *a priori* probabilities, $P(C_i)$ $\forall i$, are estimated from the proportion of each category in the training sample set,

$$P(C_i) = \frac{N_i}{N} \tag{B.10}$$

where N is the total number of objects in the sample set.

Substituting Equation B.10 into Equation B.9, the entropy score reduces to,

$$H' = -\frac{1}{N}\sum^{N}\log_2 P(C_i|\boldsymbol{x}_i) \tag{B.11}$$

[1]An entropy measure has previously been used to evaluate the performance of a Hidden Markov Model based speech recognition system (McInnes *et al.*, 1989). The entropy score presented here has been derived with the assistance of Dr. Fergus McInnes.

Thus, the entropy score, H', is the average negative logarithm (to base 2) *a posteriori* probability of an object being assigned to the correct category. The entropy is the theoretical number of bits of additional information required by any process whose task is to derive the category of the input object without error, subsequent to the Bayesian classifier. A lower entropy score indicates a Bayesian classification of better performance.

Appendix C

Publications

The following published papers by the author are directly connected with the research described in this thesis and are appended here.

1. BAGSHAW, P.C., & WILLIAMS, B.J. (1992). Criteria for labelling prosodic aspects of English speech. *Pages 859–862 of: Proc. International Conference on Spoken Language Processing*, vol. 2. Banff, Canada.

2. BAGSHAW, P.C. (1992). An investigation of acoustic events related to sentential stress and pitch accents, in English. *Pages 808–813 of: Proc. 4th. Australian International Conference on Speech Science and Technology.* Brisbane, Australia.

3. BAGSHAW, P.C., HILLER, S.M., & JACK, M.A. (1993). Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. *Pages 1003–1006 of: Proc. 3rd. European Conference on Speech Communication and Technology*, vol. 2. Berlin.

4. BAGSHAW, P.C. (1993). An investigation of acoustic events related to sentential stress and pitch accents, in English. *Speech Communication*, **13**(3–4), 333–342.[1]

[1]Reproduced with permission from Elsevier Science Publishers B.V., Amsterdam.

CRITERIA FOR LABELLING PROSODIC ASPECTS OF ENGLISH SPEECH

Paul C. Bagshaw Briony J. Williams

Centre for Speech Technology Research, University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, Scotland, UK

ABSTRACT

We report a set of labelling criteria which have been developed to label prosodic events in clear, continuous speech, and propose a scheme whereby this information can be transcribed in a machine readable format. We have chosen to annotate prosody in a syllabic domain which is synchronised with a phonemic segmentation. A procedural definition of syllables based on the grouping of phones is presented. The criteria for hand labelling the prominence of each syllable, tone-unit boundaries and the pitch movement associated with each accented syllable, are described. Work to automate this process is presented and experimental results evaluating its performance are included.

I. INTRODUCTION

The need for a large corpus of prosodically labelled English speech is motivated by the use of prosodic events in training speech synthesisers, in automated foreign language pronunciation teaching, and to aid parsers used in speech recognition to disambiguate phonetically similar, but syntactically different utterances.

Speech synthesis requires a mapping from prosodic events to a set of acoustic parameters for their realisation. Parsers and the analysis of language pronunciation, on the other hand, require the reverse mapping to provide descriptors for the acoustic correlates of prosody, and semantic and pragmatic knowledge to be extracted from these correlates. The prosodic labelling of a language corpus must therefore annotate both the linguistically significant features in speech prosody and the inflections of the acoustic parameters.

We aim to transcribe sentential stress (the prominence of syllables in continuous speech) and the pitch movement associated with any accented syllables for such systems. By initially hand labelling these prosodic aspects, a set of acoustic features are sought which will form a mapping for speech synthesis, and at the same time, enable these prosodic events to be labelled automatically given the acoustic features, for parsers and language pronunciation description. The transcription system we propose is intended to be an annotation scheme for linguistically significant prosodic events in English. It is not designed to give a detailed description of every possible inflection in an F0 contour. The set of symbols (see table 1) is designed for use by both a hand transcriber of the prosodic events and for some automated procedure.

The labelling scheme described has been used to transcribe, by hand, prosodic events in a database of 453 utterances from the English language ATR conference-registration dialogues with focus[1]. An acoustic analysis of these labels attempts to establish a correlation between a set of features chosen to characterise the acoustic parameters believed to manifest prosody, and the perceived prosodic events that are transcribed.

Continuous speech is initially segmented into phone units and labelled using a HMM-based automatic segmenter (evaluated in [18]). The phones identified are grouped into syllables. Syllable boundaries are thus synchronised with the phone boundaries. The procedure employed to group phones into syllables is described in section II. Each syllable is labelled by hand as unstressed, stressed (but not accented), stressed and accented (but not nuclear), or as the nuclear accented syllable of a tone-unit. Each syllable immediately preceding a tone-unit boundary is also marked, in order to specify the boundary location. The nuclear accented syllable of a tone-unit is (according to the "British School" of intonational phonology) the final accented syllable in that tone-unit [5]. This definition of nuclear syllables and the criteria used to determine syllable prominence are addressed in section III. Each accented syllable is associated with an additional label that describes the pitch contour movement which marked it. Thus, pitch contour labelling is also synchronised with syllable boundaries. The time location of this movement may occur before, during and subsequent to the domain of the accented syllable. Pitch contour labelling criteria are described in section IV. In section V a set of acoustic features are proposed which we intuitively feel will describe the acoustic correlates of sentential stress. These acoustic features are used to form a tree-based statistical model for a small corpus of hand labelled prosodic events. This methodology is described in section VI. Its application reveals a low correlation between the acoustic features and the events labelled, which poses questions regarding the relationship between the theory and the acoustics of sentential stress. These are discussed in section VII.

II. SYLLABIFICATION

The following procedural definition is used for syllabification.

i) Phones are grouped into syllables on a phonological rather than phonetic basis. Consonantal phones (such as [m, n, l, r, s]) which may result in schwa deletion [10, pp.297–299] [6] and take on the syllabic nucleus, are therefore syllabified as if the vowels were present. Hence, *shortest* in rapid speech is syllabified as /ˈʃ ɔ – t ᵴ t/, and *additional* as /ə – ˈd ɪ – ʃ ɳ – l/. A glottal stop that may occur before or instead of a word-final stop is treated as an instance of the underlying stop phone, and any glottalised onset to vowels is considered to be part of the vowel.

ii) Syllable boundaries are formed from the boundaries of words considered in isolation. Although in continuous speech, consonants at the end of one word can syllabify with the initial vowel of the following word [13], such resyllabification is not necessary in forming a domain in which to describe prosodic events. Thus, for example, the syllabification of *at all* differs to that of *a tall* even if /t/ is aspirated in both cases and they are phonetically identical. This approach has been adopted because the exact boundaries between syllable nuclei are not of critical im-

[1] The ATR dialogues where spoken by a female bilingual speaker of Japanese and American English.

portance, although identifying the nuclear phone is. Similarly, resyllabification is unnecessary across words that appear to blend together due to vowel deletion, as may be the case in *under a*, which is syllabified as /'ʌ n – d ɚ – ə/.

iii) The boundaries between syllables are also determined by the presence of a morphological boundary. The boundary between a free morpheme and an inflectional suffix (except -s) or a class-II derivational affix is taken to be a syllable boundary. Thus, *hopeless* is syllabified as /'h əʊ p – l ə s/ rather than /'h əʊ – p l ə s/; and *uninteresting* is syllabified as /ʌ n – 'ɪ n – t ə – r ɛ s t – ɪ ŋ/ rather than /ʌ – 'n ɪ n – t ə – r ɛ – s t ɪ ŋ/.

iv) On the basis of English phonotactics, any cluster of phones forming the onset or the coda of a syllable must also be a permissible word-initial or word-final cluster. According to this rule, *extra* may be syllabified as /'e k – s t r ə/, /'e k s – t r ə/, or /'e k s t – r ə/.

v) The 'maximal onset (and minimal coda) principle' [16] [4, pp.10–18] arbitrates between competing analyses. According to the principle, as many consonantal phones as possible form a syllable onset. Using this principle, *extra* would be syllabified as /'e k – s t r ə/. However, in cases when alternative boundaries are possible, stressed syllables tend to attract consonants more than unstressed ones, particularly in the case of ambisyllabic consonants such as [s, f] [8, pp.19–23]. When this final criterion is applied, the syllabification adopted for the example becomes /'e k s – t r ə/.

III. SENTENTIAL STRESS LABELLING

The salience of each syllable within an utterance is labelled as one of {*u*, *s*, *a*, *n*} (see table 1) on the following basis.

Sententially stressed syllables are those that are perceived as salient due to a prominence of energy and/or duration and/or pitch [7] [12, chap 4] within an utterance. The default (and therefore intonationally unmarked) pitch movement in English is a slight downwards trend in pitch [5, 11]. This movement does not give any intonational prominence to the syllable within the declination, even if that syllable is stressed on the grounds of prominent duration and/or intensity. The same situation occurs if a stressed but unaccented syllable is one in a series of gently rising pitch movements. Where there is no pitch discontinuity, there is no accent [5]. An accented syllable must also be a stressed syllable and an accompanying pitch movement must occur during the accented syllable or on a syllable before or subsequent to the perceived accented syllable [9].

Each tone-unit of an utterance will have one peak of prominence in the form of a nuclear pitch movement. The nuclear accented syllable is the syllable on which the one obligatory pitch movement occurs in a tone-unit. This is traditionally believed to be the final accented syllable in a tone-unit [15]. At present, we make use of this traditional definition.

Tone-unit boundaries are marked by placing a diacritic {:} on the label {*u*, *s*, *a*, *n*} of the syllable immediately preceding the boundary. The tone-unit boundaries are identified by two phonetic features [5, pp.204–207]. Firstly, the presence of junctural features, such as slight pauses, final lengthening and rhythmic discontinuities, can signal the end of a tone-unit. However, a pause does not necessarily correspond with a tone-unit boundary in spontaneous speech, particularly in cases of disfluency. Secondly, given that the first prominent syllable, for the majority of tone-units in an utterance, is of approximately the same pitch level [5], the boundary may be signaled by some perceivable pitch change. This change can be either a step up from a falling pitch movement, or a step down from a rising pitch movement. It may be difficult to identify such pitch resets when the tone-unit onset is low and the final accent of the previous tone-unit ends with a pitch fall, or the onset is high and follows a tone-unit whose final accent ends with a rise in pitch.

Table 1: Symbols for Sentential Stress and Pitch Movement Labelling

ASCII†	Symbol	Description
u	{*u*}	— Completely unstressed
s	{*s*}	— Stressed but unaccented
a	{*a*}	— Stressed and accented
n	{*n*}	— Nuclear accented
| *pipe*	{:}	— syllable immediately preceding a tone-unit boundary
\	{\}	— pitch accent is a fall
/	{/}	— pitch accent is a rise
v	{v}	— accent is a fall-rise
^ *hat*	{^}	— accent is a rise-fall
1	{–}	— level tone
<	{←}	— pitch movement is part of the realisation of an accented syllable to the left of this syllable
>	{→}	— pitch movement is part of the realisation of an accented syllable to the right of this syllable
- *minus*	{⁻}	— the range of the pitch movement is unusually wide (increased)
_ *underscore*	{₋}	— the range of the pitch movement is unusually narrow (decreased)
' *apostrophe*	{△}	— pitch "peak" or level tone pitch is unusually high
, *comma*	{▽}	— pitch "peak" or level tone pitch is unusually low
[	{| }	— initial part of {v} or {^} pitch movement is shallow
]	{ |}	— final part of {v} or {^} pitch movement is shallow

†The ASCII characters listed are the prosodic labels used in machine readable data.

IV. PITCH MOVEMENT LABELLING

The pitch contour of an utterance is labelled as a series of pitch movements at (or near) each accented syllable. A pitch movement is either a continuous pitch glide, for example over a long vocalic section of speech, or a discrete pitch jump from one level to another over a series of syllables. Each pitch movement in an utterance is labelled as one of the five categories {\, /, v, ^, –} (see table 1).

A description is associated with each and every syllable labelled as accented (or nuclear accented) to mark the direction of pitch movement on this and any following unaccented syllables. These labels should only be time aligned with an unstressed or an accented (nuclear or otherwise) syllable {*u*, *a*, *n*}, but not with a stressed (but unaccented) syllable {*s*}. (Any stressed syllable corresponding with a time aligned pitch movement label should be marked as an accented syllable.) If the pitch movement is aligned with an unstressed syllable {*u*}, a diacritic is applied to the pitch movement label in order to indicate whether the pitch movement is part of the realisation of the nearest accented syllable {*a*, *n*} to the left {←} or the nearest one to the right {→}. There may be more than one pitch movement associated with an accented syllable; for example, if there is a rise-fall pitch movement in the realisation of an accented syllable but the rise occurs on a preceding unstressed syllable and the fall occurs on a succeeding unstressed syllable. The uses of these diacritics enable the inflections of the F0 contour to be described while maintaining a

transcription of the perceived pitch movement.

Pitch range markings are used to describe the extent of the movement in a pitch glide and the distance between levels of a pitch jump, but not for level tone. If the pitch range is distinctively wider or narrower than expected for a particular contrastive effect, it is marked with a diacritic {ˉ, _} on the pitch direction labels. Diacritics are also applied to these labels if the "peak" part of a pitch movement (the initial part of a fall {\}, the final part of a rise {/}, and the mid-section of a fall-rise or rise-fall {∨, ∧}) or the pitch of a level tone {−} is unusually high {△} or low {▽} for the particular speaker. In order to describe occurrences of pitch fall-rise and rise-fall with a particularly shallow rise or shallow fall, two further diacritics are included. These are used to represent, for example, fall-shallow rise as {∨}.

V. ACOUSTIC FEATURES

A set of acoustic features must be extracted from the raw speech waveform in order to automatically identify syllable prominence and pitch movements. In our preliminary stages of producing an automatic prosodic labelling algorithm, eighteen features are used to describe what are we believe to be the acoustic correlates of stress (duration, intensity and fundamental frequency).

The energy and fundamental frequency of the speech waveform (sampled at 20kHz) are measured for 20ms frames of speech at 5ms intervals so that values are synchronised with the cepstral coefficients and lower three formant frequencies used in the auto-segmentation process. The fundamental frequency (F0) is determined using a slightly enhanced version of the pitch tracker described in [14]. In order to measure the signal energy, each frame is passed through a Blackman-Harris window and an amplitude spectrum is calculated using a 512-point FFT. The total energy for the frequency range of 50Hz–2kHz is determined by summation of the corresponding frequency bins. Each frame energy value is then expressed in decibels with respect to the maximum frame energy for the utterance. This process forms an utterance-normalised sonorant energy contour. Both the raw F0 contour and the energy contour are smoothed using a 3-point non-linear median filter and a 5-point hanning window [17].

The phone given by auto-segmentation which forms the nucleus of a given syllable is identified by the following procedure. The phones in the syllable are split into two groups on the basis of whether or not they are a member of the set of vocalic phones and potentially syllabic consonantal phones (currently, all vowels plus [l, m, n, r]). Each phone is associated with the maximum sonorant energy within its tenure. If there are phones in the syllable which are members of this set, then the one whose associated energy is greatest, is selected as the syllable nucleus. Otherwise, none of the syllable phones are [vowel, l, m, n, r] and the phone with the greatest maximum sonorant energy is selected. The duration associated with any syllable in determining its prominence is the duration of its nuclear phone – this will be referred to as the "syllable duration". Using the duration of the entire syllable or the duration of all consecutive sonorants in the syllable as this measure has not yet been investigated.

Each syllable in an utterance is characterised by the maximum soronant energy within its tenure (syllable energy), its "syllable duration", the maximum F0 value within its tenure, the F0 values at the beginning and at the end of the syllable, and an F0 slope in Hz per second which describes the rate of change in F0 through any voiced regions of the syllable. The syllable energy and "syllable duration" are Z-score normalised to eliminate phone-specific effects [2]. For each phone type, the mean and population standard deviation of the syllable energy/duration is determined. Then, for each token of that phone type, the syllable energy/duration is normalised by subtracting the mean and dividing by the population standard deviation. Hence, for each syllable, there are six acoustic features extracted – phone-normalised duration, phone-normalised energy, maximum F0, start-time F0, stop-time F0, and F0 slope. In automatically establishing the prominence of any syllable in an utterance, these six features for the current, previous and next syllable are used, giving a total of eighteen features per syllable.

The F0 features are also normalised so that each movement is independent of its absolute F0 values. Our intuition suggests that F0 change is the significant factor, not the absolute F0 values. Normalisation of the nine F0 parameters (the maximum F0, start-time F0, and stop-time F0 for the current, previous and subsequent syllables), is performed by determining the minimum value of these parameters and subtracting it from each. The change in F0 through the syllables is therefore described independently of the absolute height of the F0 movement.

Table 2: Confusion Matrix of Sentential Stress Labelling by Hand and by Automation – cyclic exclusion of each utterance during training

		Automatic Label			
		a,n	*s*	*u*	total
Hand Label	*a,n*	889 (12.3%)	72 (1.0%)	849 (11.7%)	1810 (25.0%)
	s	237 (3.3%)	72 (1.0%)	673 (9.3%)	982 (13.6%)
	u	567 (7.8%)	142 (2.0%)	3731 (51.6%)	4440 (61.4%)
	total	1693 (23.4%)	286 (4.0%)	5253 (72.6%)	7232 (100.0%)

Misclassification error rate = 2540/7232 (35.1%)

Table 3: Confusion Matrix of Sentential Stress Labelling by Hand and by Automation – all utterances used during training

		Automatic Label			
		a,n	*s*	*u*	total
Hand Label	*a,n*	1143 (15.8%)	44 (0.6%)	623 (8.6%)	1810 (25.0%)
	s	240 (3.3%)	113 (1.6%)	629 (8.7%)	982 (13.6%)
	u	334 (4.6%)	51 (0.7%)	4055 (56.1%)	4440 (61.4%)
	total	1717 (23.7%)	208 (2.9%)	5307 (73.4%)	7232 (100.0%)

Misclassification error rate = 1921/7232 (26.6%)

VI. APPLICATION OF A TREE-BASED STATISTICAL MODEL

The sentential stress and pitch movements associated with accented syllables have been hand labelled in the ATR database of 453 utterances using the symbols given in table 1. The prosodic transcription was done by only one labeller.

The automatic prosodic labelling algorithm is still in its infancy and so the acoustic features described in section V are being used only to identify any given syllable in an utterances as either unstressed, stressed or accented (nuclear or otherwise). Distinguishing pitch movement types has not yet been incorporated.

The acoustic features are used as parameters to a tree-based statistical model (using "S" [3]). The model is trained on all but one of the utterances in the database. The tree classifies each hand-transcribed sentential stress label on the basis of the

features given. This tree is then used to predict the labels for the utterance that was not included in the training set. These automatically generated labels are compared with those given by hand. This process is repeated in a cyclic fashion for all the utterances and the comparisons are summed. The confusion matrix (table 2) indicates the number of occurrences that each hand-transcribed label is predicted as accented {*a*, *n*}, stressed {*s*} or unstressed {*u*} using this process.

In order to give an indication of the dependency of the automatic labels on the method used, table 3 shows a similar confusion matrix generated when the test utterance is included in the training data.

VII. DISCUSSION

The misclassification error rate of 26.6% is quite promising given that the selection of the acoustic features that have been used is based on intuition. This, however, may not be the only contributing factor to erroneous classifications. It could be that the acoustic features are in fact closely related to the prosodic events labelled, but that the tree-based statistical model is not the most appropriate method to classify these events given the acoustic features (this is supported by the considerable difference between tables 2 & 3). Alternatively, the acoustic features presented could be insufficient to characterise the prosodic events. For example, it is likely that representing F0 movements across a three-syllable window is restrictive, given that such movements can clearly span many or part of syllables. It may be that the labelling scheme is an inadequate system for describing sentential stress and the pitch movements as perceived by the transcriber. This can be illustrated by the fact that sentential stress is not a simple binary distinction between stressed and unstressed. In ambiguous cases, the transcriber uses linguistic knowledge not evident in the acoustics. For example, the syllable in question will be marked as sententially stressed only if it can be lexically stressed. This may lead to every occurrence of schwa being marked as unstressed regardless of the acoustic evidence. With such linguistic knowledge unavailable to the tree-based model, confusions will inevitably arise between the hand labels and automatic labels.

It is most likely that the classification errors are due to some combination of all these factors, although the extent to which any one factor effects the error rate is difficult to determine. The correct-classification rate of 73.4% is, however, close to the percentage of correlating labels between two hand labellers – in the prosodic labelling of the Lancaster/IBM spoken English corpus, transcribers achieved 72% agreement for seven categories of sentential stress labels {\, /, ∨, ∧, –, *s*, *u*} and 83% agreement for the categories "accented"/ "stressed"/ "unstressed" [1].

ACKNOWLEDGEMENTS

Thanks to Keith Edwards, Sally Bates, Alex Monaghan, Nick Campbell, Jim Hieronymus, and Bob Ladd for their valuable assistance. This work has been supported by ATR Interpreting Telephony Research Laboratories, Kyoto, Japan.

References

[1] P. Alderson and G. Knowles. *Working with speech.* Longman, London, in press.

[2] W.N. Campbell. Evidence for a syllable-based model of speech timing. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 9–12, Kobe, Japan, 1990.

[3] L.A. Clark and D. Pregibon. Tree-based models. In J.M. Chambers and T.J. Hastie, editors, *Statistical Models in S*, chapter 9, pages 377–419. Wadsworth & Brooks, Pacific Grove, California, 1992.

[4] E. Couper-Kuhlen. *An Introduction to English Prosody.* Edward Arnold (Publishers) Ltd., London, 1986.

[5] D. Crystal. *Prosodic Systems and Intonation in English.* Cambridge University Press, Cambridge, U.K., 1969.

[6] J.M. Dalby. *Phonetic Structure of Fast Speech in American English.* PhD dissertation, Indiana University Linguistics Club, Bloomington, Indiana, 1986.

[7] D.B. Fry. Duration and intensity as physics correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27(4):765–768, 1955.

[8] E.C. Fudge. *English Word-Stress.* George Allen & Unwin, London, 1984.

[9] E. Gårding and Gerstman. The effect of changes in the location of an intonation peak on sentence stress. *Studia Linguistica*, 14:57–59, 1960.

[10] A.C. Gimson. *An Introduction to the Pronunciation of English.* Edward Arnold, London, second edition, 1970.

[11] D.R. Ladd. Peak features and overall slope. In A. Cutler and D.R. Ladd, editors, *Prosody: Models and Measurements*, chapter 4, pages 39–52. Springer-Verlag, Heidelberg, Germany, 1983.

[12] I. Lehiste. *Suprasegmentals.* The Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1970.

[13] I. Maddieson. Phonetic cues to syllabification. In V.A. Fromkin, editor, *Phonetic Linguistics (essays in honor of P.Ladefoged).* Academic Press Inc., London, 1985.

[14] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, ASSP–39(1):40–48, 1991.

[15] J.D. O'Connor and G.R. Arnold. *Intonation of Colloquial English.* Longman, London, second edition, 1973.

[16] E. Pulgram. *Syllable, Word, Nexus, Cursus.* Mouton, The Hague, 1970.

[17] L.R. Rabiner, M.R. Sambur, and C.E. Schmidt. Applications of non-linear smoothing algorithms to speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP–23(6):552–557, 1975.

[18] M.S. Schmidt and G.S. Watson. The evaluation and optimization of automatic speech segmentation. In *Proc. 2nd. European Conference on Speech Communication and Technology*, volume 2, pages 701–704, Genova, Italy, 1991.

AN INVESTIGATION OF ACOUSTIC EVENTS RELATED TO SENTENTIAL STRESS AND PITCH ACCENTS, IN ENGLISH

Paul C. Bagshaw
ATR Interpreting Telephony Research Laboratories
(Visiting from the Centre of Speech Technology Research, University of Edinburgh)

ABSTRACT - An algorithm is described to abstract acoustic parameters of a speech waveform to give a scalar measure of the relative stress and pitch movement of each group of phones which can consist of a single prominence. A method of identify such groups using acoustic information is given. The abstracted parameters are used to locate sentential stress and pitch accents in English speech. These are compared with a hand-labelled prosodic transcription.

I. INTRODUCTION

We wish to label prosodic events in English speech. Prosodic events marked by hand vary considerably from labeller to labeller and may be marked inconsistently within any labellers transcription. (Pickering, et.al, in press) show that two transcribers select the same prosodic label (level, fall, rise, fall-rise, rise-fall, stressed but unaccented or unstressed) for 72% of syllables. This paper presents a method of automaticly transcribing prosodic events with the relative stress of any syllable and the extent of pitch movements being described as a scalar rather than as a discrete level. The method involves a series of abstractions of acoustic parameters which aims to isolate the prosodic variations in duration, energy and fundamental frequency from the microprosodic variations.

The grouping of phones into syllables which can constitute at most a single prominence in the utterance is used as a domain for transcribing prosodic events. The method used to produce the phone groups is described in section II. The prosodic content of each phone group is described by giving it a measure of its relative stress in the utterance and, if the group is accented, the type of pitch movement (level, fall, rise, fall-rise or rise-fall) and the relative extent of the movement. In deducing the relative stress, the prosodic variations of duration, energy and fundamental frequency (F0) are abstracted from the speech waveform acoustics. The formation of a piece-wise F0 contour to remove microprosodic variations is described in section III. The piece-wise units crossing each syllable are abstracted into one of five types of pitch movement. Each syllable is then marked as either prominent (sententially stressed) or not prominent (sententially unstressed), and if it is found to be prominent and pitch salient, it is marked as accented (section IV). These markings are compared with those transcribed by hand.

II. SYLLABIFICATION FROM ACOUSTIC PARAMETERS

An algorithm is described to group phones given by an automatic phonemic segmentation system into syllable sized items based on sonorant energy. The syllabification of speech from acoustic parameters groups phones according to the manner in which the speaker formed the utterance rather than that dictated by a set of phonological rules. The syllabification makes full use of the phone boundary and label information given by auto-segmentation and uses the sonorant energy contour of the utterance to determine their grouping.

The energy contour for a speech waveform (sampled at 20kHz) is calculated from 20ms frames at 5ms intervals so that values are synchronised with the cepstral coefficients and lower three formant frequencies used in the auto-segmentation process. Each frame is passed through a Blackman-Harris window and the frequency bins of an amplitude spectrum (512-point FFT) corresponding to the range 50Hz–2kHz are accumulated. These energy values are expressed in decibels with respect to the maximum frame energy in the utterance to form an utterance-normalised sonorant energy contour. The contour is processed by a three-frame median filter and five-frame hanning window smoother (Rabiner, et.al, 1975) in order to remove small perturbations which arise during frames of speech with low fundamental frequency (typically less than two pitch periods per analysis frame).

Proc. 4th. Australian International Conf. on Speech Science & Technology, 808–813.

All minima in the energy contour are located and form candidates for syllable boundaries. The areas of silence identified by the auto-segmentation are respected and so the minima within the tenure of such areas are believed to be due to variations in background noise. Each boundary between a silence and a phone label is taken as either the beginning or the end of a syllable. The nearest candidate to such a boundary is therefore moved to align with it and all those residing within the silence section are disregarded. The regions between all the remaining energy minima are taken to be potential syllables with a start time given by the nearest left-hand-side minimum's location, and the nearest right-hand-side minimum's location giving the stop time. It is determined if the location of each of these potential syllables overlaps more than 50% of any auto-segmented vowel. If it overlaps more than one vowel segment in this way, then the vowel segment with the maximum sonorant energy is taken to be the nucleus of the syllable. If no such overlap occurs, then it is determined if the location of the potential syllable overlaps more than 50% of one of the possible syllabic consonant segments /l, m, n, r/. Again, if it overlaps more than one of these, the one with the maximum sonorant energy is selected as the syllabic nucleus. If there is insufficient overlap, the region between the minima does not correspond to a syllable unit and either the l.h.s. or r.h.s. minimum is disregarded as a syllable boundary candidate — whichever has the highest energy and does not correspond to a phone/silence boundary. The newly formed region is then taken to be a potential syllable and the process is repeated. The resultant syllabification has boundaries located at positions of minimum sonorant energy in the utterance. These boundaries may be aligned with the auto-segmentation by moving each syllable boundary to the nearest phone boundary.

A database of 453 utterances from the English language ATR conference-registration dialogues has be syllabified using the above algorithm and by using a phonologically based syllabification (Bagshaw & Williams, 1992). There is a large correlation between the two resultant syllabic domains (see table 1). The missing syllable boundaries are due to the occurrences of vowel/vowel boundaries for which there is no valley in the sonorant energy between them. When this case arises, often one of the vowels is a schwa; for example, the phonological syllabification of "my address" as /m aɪ – ə – d r ɛ s/ can be grouped on an acoustic basis as /m aɪ ə – d r ɛ s/. Conversely, extra syllable boundaries occur when the sonorant energy dips within the tenure of the phonologically based syllable at a vowel/vowel boundary or vowel/syllabic consonant boundary; for example the phones in "tour" /t ʊ ɚ/ can be grouped as /t ʊ – ɚ/ on an acoustical basis, and for the word "forms" /f ɔ r m s/ phones are grouped as /f ɔ – r m̩ s/ as its pronunciation tends towards that of "forums".

Syllabification using acoustic parameters in this manner clusters phones with a vowel or syllabic consonant as its nucleus and containing a single burst of sonorant energy. The duration of the nucleus and the maximum sonorant energy within it are used in determining its relative prominence. The duration and energy variations are mainly attributed to phone type. These parameters are therefore Z-score normalised with respect to the phone type of the nucleus in order to compensate for segmental variations. (Campbell, 1990) uses a similar normalisation but on a phone by phone basis rather than on the basis of syllable nuclei. The mean duration and maximum energy and their population standard deviations are determined for each phone type from a training database of 200 phonemically balanced utterances. The Z-score normalisation of a nuclear phone's duration or intensity simply involves subtracting the mean value and dividing by the population standard deviation for that phone type.

An example of these processes is shown in figure 1. Part (a) shows the speech waveform and its corresponding automatic segmentation using MRPA labels (Edinburgh University's machine-readable phonemic alphabet). Part (b) gives the utterance normalised sonorant energy contour and transcription-aligned syllable boundaries with a MRPA label indicating the phone forming the nucleus. The Z-score normalised duration and energy for each syllable nucleus is given in part (c). These will be discussed further in section IV.

III. THE FORMATION OF A PIECE-WISE F0 CONTOUR

A fundamental frequency (F0) contour produced by a pitch determination algorithm (PDA) can be expected to contain values which are inaccurate, such as instances of pitch octave errors. Any PDA will also make erroneous classifications of sections of speech as voiced or unvoiced. A personal evaluation of a slightly enhanced version of the PDA described in (Medan, et.al, 1991) (which is used in this study) has been found to estimate F0 with consistently less than 1% gross pitch errors and less than 16% of speech classified as voiced or unvoiced incorrectly, when compared with F0 determined from

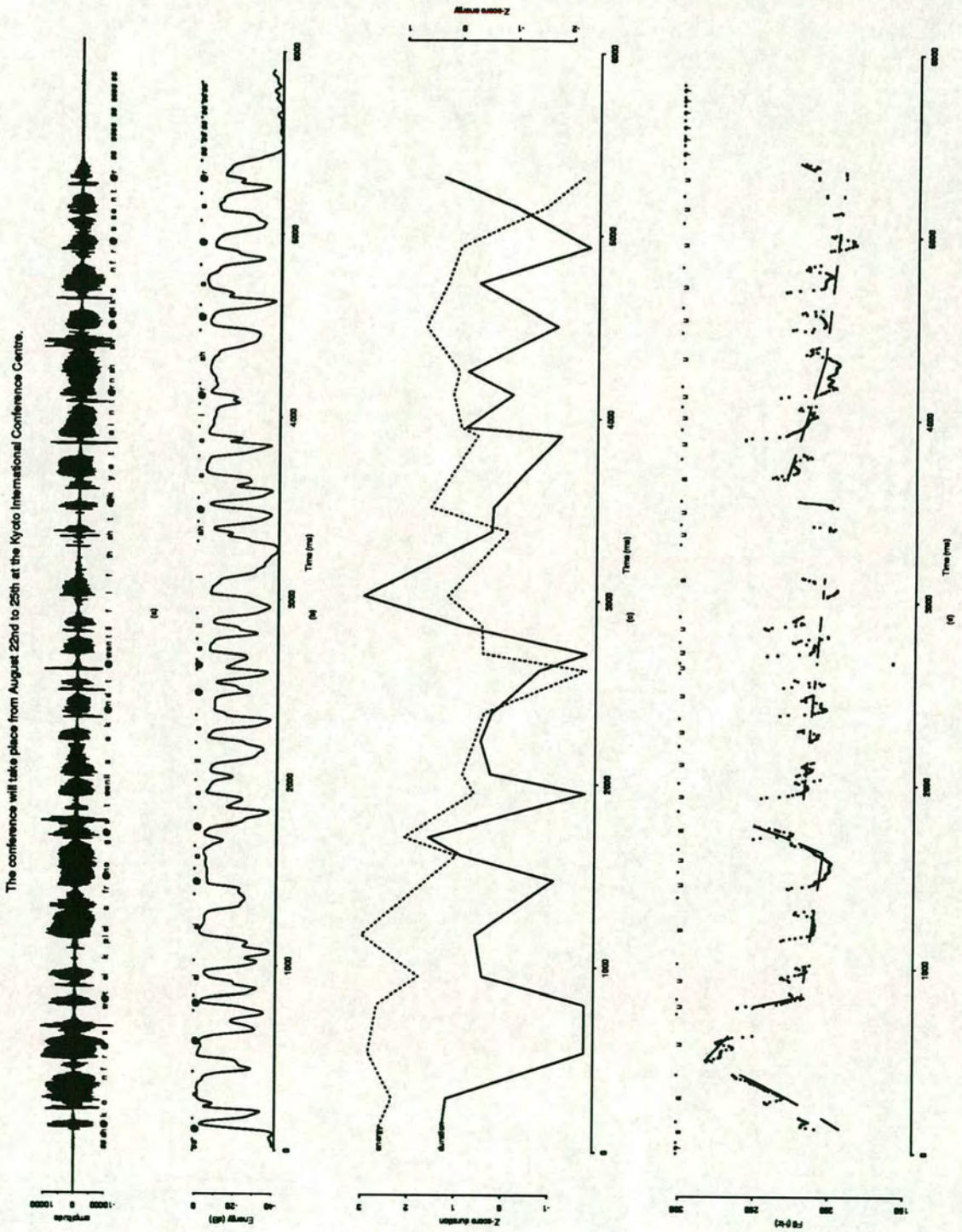


Figure 1: Example of the abstraction of acoustic features related to prosodic events

Proc. 4th. Australian International Conf. on Speech Science & Technology, 808–813.

Table 1: Comparison of Phonologically Based and Acoustically Based Syllabifications

Number of syllables				
from a phonological basis	from the acoustics	Match	Missing	Extra
7299 (100.0%)	7011 (96.1%)	6980 (95.6%)	-319 (4.4%)	+31 (0.4%)

laryngograph data. In order to eliminate the majority of octave errors and reduce microprosodic perturbations, the contour is initially processed by a three-frame median filter and three-frame hanning window smoother (Rabiner, et.al, 1975). The frames of speech analysed by the PDA are in synchronisation with those used in calculating the sonorant energy contour and in the auto-segmentation. The resultant contour is an excellent estimate of the fundamental frequency of the speech waveform, but it does not form a descriptor of utterance intonation alone as microprosodic variations are also present. A process of piece-wise linear stylisation of the contour aims to eliminate such variations.

The algorithm used to perform the stylisation is based on the technique described by (Scheffers, 1988) and incorporates the robust least median of squared residuals regression (LMedR) (Rousseeuw & Leroy, 1987). The F0 values describing the contour (excluding values which equal zero to represent unvoiced speech) are converted to the semitone scale using the relationship $F0_{semitone} = 12\log_2(F0_{hertz}/55)$. Significant turning-points in the F0 contour are located, these points are modified to prevent contour discontinuities other than at the boundaries between unvoiced and voiced speech, and a new contour is generated by interpolating between them.

The following process is used to identify the turning-points. Starting with the first voiced frame, LMedR analysis is applied to a window of w frames corresponding to voiced speech, where w is initially set to 5. The final frame in this window is taken to be a turning-point candidate. The F0 value of the subsequent frame is predicted using the coefficients of the LMedR analysis. If the absolute difference between the actual and predicted F0 values is less than or equal to some level of permitted variation in F0 (1 semitone), then the candidate is not a turning-point, the window length w is incremented to include the next voiced frame, and the above process is repeated. The repetition of this process terminates when the turning-point candidate is the final voiced frame in the F0 contour. Otherwise, when the absolute difference is greater than the permitted F0 variation, either this subsequent F0 value constitutes some type of irregularity in the F0 contour or the candidate could be a true turning-point. To determine which is the case, the F0 value of the next voiced frame is also predicted. If the absolute difference between the predicted and actual values is once again greater than the permitted F0 variation, and this situation arises for all following frames up to either the final voiced frame in the contour or such that the duration of this discontinuity is greater than some minimum permitted level (100ms), which ever occurs first, then the candidate is said to be a true turning-point. Otherwise, the length of the window w is increased to include the first frame for which the absolute difference in actual and predicted F0 values was less then or equal to the permitted variation, but not those for which it was greater, and the LMedR analysis process is repeated. If the candidate was found to be a turning-point and if it corresponds to a voiced frame immediately preceding a frame of unvoiced speech, then the first frame of the next voiced region is also designated as a turning point. This process is then repeated with the length of the window w reset to 5 and the first frame of the window is set to the frame of the most recent turning-point found. The first and final voiced frames of the non-stylised contour are also assigned as turning-points.

In order to ensure that discontinuities in the stylised F0 contour only occur at unvoiced sections of speech, the fundamental frequency at each turning-point of the new contour is determined in a way which depends upon the voicing state of the frames adjacent to it. For any given turning-point (tp) at frame f_{tp} with original fundamental frequency $F0_{tp}$, the LMedR coefficients s_{tp} (slope) and i_{tp} (intercept) of the windowed points preceding the turning-point are known. The modified fundamental frequency $F0'_{tp}$ is given as,

$$F0'_{tp} = \begin{cases} 0.5(s_{tp}.f_{tp} + i_{tp} + s_{tp+1}.f_{tp} + i_{tp+1}) & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ voiced} \\ s_{tp+1}.f_{tp} + i_{tp+1} & \text{if frame } f_{tp}-1 \text{ unvoiced \& frame } f_{tp}+1 \text{ voiced} \\ s_{tp}.f_{tp} + i_{tp} & \text{if frame } f_{tp}-1 \text{ voiced \& frame } f_{tp}+1 \text{ unvoiced} \\ F0_{tp} & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ unvoiced} \end{cases} \quad (1)$$

The new stylised contour is then created by linear interpolation of F0 between each turning-point (f_{tp}, $F0'_{tp}$) and by reseting each frame that is unvoiced in the non-stylised contour to an unvoiced state in

the new one. The resultant data is then coverted back to a Hertz scale. An example of this piece-wise stylisation is shown in figure 1(d).

The F0 contours produced for the database of 453 utterances have been stylised using this method. Cepstral resynthesis of the speech from both the original F0 and the piece-wise F0 have been compared by ear. Of these, I felt that 405 (89.4%) contain no perceptual difference in prosodic content.

IV. PROSODIC ABSTRACTION

The piece-wise F0 contour will, for some utterances, contain small units which are erroneous, ie. do not correspond to part of pitch movements. Only those piece-wise units which, at some time, run through any part of a syllable nuclear phone (where F0 estimation is expected to be reliable) are treated as being part of a pitch movement. Moreover, the absolute F0 range of a piece-wise unit is not of interest as it will vary from speaker to speaker, but its extent relative to other units in the utterance is. The relative extent of each piece-wise unit is calculated by first locating a regression line which best fits the contour turning points using LMedR analysis. A by-product of the LMedR is the standard deviation, $\sigma_{l\,MedR}$ of the points from the resultant linear model. The absolute F0 at each turning point is then converted by subtracting its modelled value and dividing by the standard deviation, $\sigma_{l\,MedR}$. This effectively compensates for any long term declinative tendency that may be exhibited in the fundamental frequency contour, and expresses the F0 values relative to an utterance dependent datum.

Once the relative extent of each piece-wise unit has been established, they are combined to form pitch movement descriptors. The pitch movements facilitated are level, fall, rise, fall-rise and rise-fall $\{-, \backslash, /, \vee, \wedge\}$, as given by the "British School" of intonational phonology (Crystal, 1969). Each piece-wise unit crossing any part of a syllable nuclear phone is classified as either level, fall, or rise. Let $F0_{start}$ be the relative F0 height at the start of the piece-wise unit and that at the end of the unit be $F0_{end}$. The piece-wise unit is classified on the following basis,

$$\text{pitch movement} = \begin{cases} \backslash & \text{if } F0_{start} - F0_{end} > 0.75\sigma_{l\,MedR} \\ / & \text{if } F0_{start} - F0_{end} < -0.75\sigma_{l\,MedR} \\ - & \text{otherwise} \end{cases} \quad (2)$$

When more than one piece-wise unit crosses any particular nucleus, they are combined by initially taking all adjacent units with the same pitch movement classification and joining them into one. A join is made by setting $F0_{start}$ to that of the first unit, $F0_{end}$ to that of the second unit, and reclassifying using equation 2. In the database of 453 utterances, consisting of 7299 syllables, there were only 4 syllables for which more than two units remained after this process. These all contained some error which originated in the F0 estimation and are ignored. If there are two remaining units (their classifications must differ), and if either is classified as level $\{-\}$, then they too are joined in the same way. Otherwise, one is a fall $\{\backslash\}$ and the other is a rise $\{/\}$. These are combined to give a single movement classified as either a fall-rise $\{\vee\}$ or rise-fall $\{\wedge\}$ depending on their order, and the relative level at their mid point is kept. Thus, for the fall-rise and rise-fall classifications, the extent of both the onset and coda of the movement are known.

Having established the shape of the pitch movement of each syllable in this way, and with knowledge of the Z-score normalised syllable nucleus duration and intensity measures, we determine if any given syllable is prominent in the utterance (sententially stressed $\{s\}$) and if it is pitch salient (accented $\{a\}$). A syllable is marked as prominent if both its normalised duration and its normalised energy are greater than that of both its nearest neighbours, and if both are greater than 0.75 standard deviations from the mean value. (The value 0.75 is arbitrary.) It is similarly marked if either the normalised duration or normalised energy is the maximum for the utterances. A syllable is marked as accented using a decision filter three pitch movements wide (Hieronymus, 1989).

An example of such prominence detection is illustrated at the top of figure 1(d). The database of 453 utterances has been automatically prosodically marked in this way and compared with those transcribed by hand (see table 2). The transcriptions are equal for 61.6% of the syllables. Of the unstressed $\{u\}$ hand labels marked as either accented or stressed automatically, 216 were syllables with a schwa nucleus. This indicates that the hand transcriber may be marking syllables as sententially stressed only if they can be lexically stressed. There is also a noticeably large number of syllables labelled by hand as accented or stressed that are marked as automatically as unstressed, indicating that the hand labeller

Table 2: Confusion Matrix of Prosodic Transcription by Hand and by Automation

		Automatic Label			
		a	s	u	total
Hand Label	a	566 (7.8%)	177 (2.4%)	1075 (14.7%)	1818 (24.9%)
	s	128 (1.8%)	71 (1.0%)	792 (10.9%)	991 (13.6%)
	u	404 (5.5%)	226 (3.1%)	3860 (52.9%)	4490 (61.5%)
	total	1098 (15.0%)	474 (6.5%)	5727 (78.5%)	7299 (100.0%)

Correct classification rate = 4497/7299 (61.6%)

may be using acoustic parameters other than those described previously in this paper. For example, syllables whose nucleus is "fully articulated" are often marked as stressed by hand. Such measures are currently unavailable to the automatic prosodic transcription algorithm.

V. CONCLUSION

An algorithm to group phones into syllables which can consist of only one prominence has be described. This forms a domain in which to transcribe prosodic events. The domain correlates closely with a phonologically based syllabification. The piece-wise stylisation of a fundamental frequency contour to eliminate micro-prosodic variations in F0 has been further abstracted to form pitch movements for each syllable. The extend of these movements are known relative to other pitch movements in the utterance. The prominence of each syllable has been determined from these parameters and compared with a hand-labelled prosodic transcription. Although the correlation between labels (61.6%) is lower than one would hope, the reason for this may not be because the algorithm performs "poorly" but because it appears that the hand labels transcribe aspects of speech that are not apparent in the waveform acoustics used.

ACKNOWLEDGEMENTS

Thanks to Nick Campbell, Jacqueline Vaissière, Jim Hieronymus and Peter Meer for their valuable assistance and suggestions.

REFERENCES

Bagshaw, P.C. & Williams, B.J. (1992) *Criteria for labelling prosodic aspects of English speech*, In Proc. International Conference on Spoken Language Processing, Banff, Canada (forthcoming).

Campbell, W.N. (1990) *Evidence for a syllable-based model of speech timing*, In Proc. International Conference on Spoken Language Processing, Kobe, Japan, vol.1, 9–12.

Crystal, D. (1969) *Prosodic Systems and Intonation in English*, (Cambridge University Press: Cambridge).

Hieronymus, J.L. (1989) *Automatic sentential vowel stress labelling*, In Proc. European Conference on Speech Communication and Technology (EUROSPEECH-89), Paris, vol.1, 226–229.

Medan, Y., Yair, E. & Chazan. D. (1991) *Super resolution pitch determination of speech signals*, IEEE Trans. Signal Processing, ASSP–39(1), 40–48.

Pickering, B., Williams, B. & Knowles, G. (in press) *Analysis of transcriber differences in the SEC*, In Alderson, P. & Knowles, G. (eds.) Working with Speech, chapter 4, (Longman: London).

Rabiner, L.R., Sambur, M.R. & Schmidt, C.E. (1975) *Applications of non-linear smoothing algorithms to speech processing*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP–23(6), 552–557.

Rousseeuw, P.J. & Leroy A.M. (1987) *Robust Regression and Outlier Detection*, (Wiley: New York).

Scheffers, M.T.M. (1988) *Automatic stylization of F0-contours*, In Ainsworth, W.A. & Holmes, J.N. (eds.) Proc. of 7th. FASE Symposium, Edinburgh, vol.3, 981–987.

ENHANCED PITCH TRACKING AND THE PROCESSING OF F0 CONTOURS FOR COMPUTER AIDED INTONATION TEACHING

P.C. Bagshaw, S.M. Hiller, and M.A. Jack

Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh, EH1 1HN, Scotland, UK

ABSTRACT

A comparative evaluation of several pitch determination algorithms (PDAs) is presented. Fundamental frequency estimates, F0, are compared with laryngeal frequency estimates, Lx. An algorithm is presented which enables Lx contours to be generated from laryngograph data. We seek the most accurate method of F0 extraction in order to minimise errors propagating into subsequent prosodic analysis. The super resolution pitch determinator [3] performs well relative to the other PDAs studied. Modifications made to this algorithm are described, which radically reduce the number of gross F0 errors and improve the classification of voiced and unvoiced sections of speech. The raw F0 contours produced by this enhanced algorithm are processed to form schematised contours used in computer aided intonation teaching. The series of processes used in the schematisation is described.

Keywords: *Pitch tracking, Intonation, Language teaching*

1 INTRODUCTION

The fundamental frequency of speech plays an important role in the prosodic features of stress, rhythm, and intonation. The understanding and appropriate use of prosody is an important component of foreign language learning, for both comprehension and intelligibility. Computer aided teaching of intonation therefore requires the determination of *F*0 as an initial process in the automated assessment of the speech of a non-native student. Determining *F*0 is not a simple task, and many approaches have been reported [2]. The selection of PDAs investigated here covers a range of techniques which use both time domain and frequency domain representations of speech. An evaluation of the algorithms, based on the use of laryngograph data, is described in Section 2. A method of forming a 'reference' contour from laryngograph data is presented. *F*0 contours generated by each PDA are compared with the 'reference' *Lx* contours. The evaluation shows that the super resolution pitch determinator has the potential to form accurate *F*0 contours from low-pass filtered speech. Errors occurring during *F*0 extraction must be minimised to prevent them from propagating into the prosodic analysis. Enhancements described in Section 3 are made to this algorithm in order to minimise *F*0 errors. The *F*0 contour produced by a PDA is, however, not manipulated solely by linguistic and paralinguistic effects. An *F*0 contour is also affected by segmental content, micro-perturbations, the speaker's anatomy and physiology, and errors involved in its determination from the speech waveform. Such *F*0 variations need to be removed in order to facilitate the comparison of a student's intonation with that of a native speaker. This is performed by a series of post-processes which schematises a raw *F*0 contour and is described in Section 4.

2 EVALUATION OF PITCH DETERMINATION ALGORITHMS

Seven PDAs are investigated. Their selection was influenced by availability, by ease of implementation, and by the desire to examine methods of *F*0 extraction which use radically different techniques.

- Cepstrum pitch determination (CPD) [4].
- Feature-based pitch tracker (FBPT) [6].
- Harmonic product spectrum (HPS) [11] [5].
- Integrated pitch tracking algorithm (IPTA) [12].
- Parallel processing method (PP) [1].
- Super resolution pitch determinator (SRPD) [3].
- Enhanced version of SRPD (*e*SRPD).

The functionality of all of the algorithms is dependent upon certain thresholds and pre-determined parameters, some of which are common across algorithms. In order to set a degree of similarity between the PDAs, all are required to present a computed *F*0 value at 6.4ms intervals. The values are limited to the ranges of 50Hz–250Hz for male speakers and 120Hz–400Hz for female speakers. In cases where a fixed-length analysis frame is required by an algorithm, the frame duration is set to 38.4ms. This duration enables at least two signal periods to reside within the frame for all *F*0 values greater than 52Hz, and allows sufficient data for cepstral and spectral analysis techniques. The speech data is sampled at 20kHz using a 16-bit analogue-to-digital converter. Some of the PDAs require a low-pass filtered version of this data, which is produced by an FIR filter with a -3dB cut-off at 600Hz

and rejection greater than -85dB above 700Hz.

2.1 A Laryngeal Frequency Tracker

The 'reference' contours are created from laryngograph data recorded simultaneously with speech by using a simple 'pulse' (Fig. 1) location algorithm and deriving the duration between successive pulses.

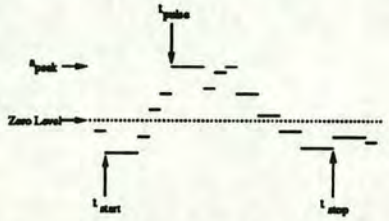


Fig. 1: Laryngograph 'Pulse'

The pulse start time t_{start} is the first sample for which the amplitude is less than zero and less than or equal to the amplitude of following samples. The pulse stop time t_{stop} is the last sample for which the amplitude is less than zero and less than or equal to the amplitude of preceding samples. The pulse width t_{width} is defined as the difference between t_{start} and t_{stop}. The pulse peak amplitude a_{peak} is the maximum amplitude of samples between t_{start} and t_{stop} (always greater than zero.) The pulse instant t_{pulse} is defined as the time of the first of these samples with an amplitude a_{peak}. For laryngograph data sampled at 20kHz, a pulse at t_{pulse} is classed as a marker of the glottal closure instant if the pulse width t_{width} is greater than four samples and the pulse peak amplitude a_{peak} is greater than some arbitrary threshold value. The duration between one pulse instant t^n_{pulse} and the next t^{n+1}_{pulse} is calculated and converted to Hertz. If the value lies within a limited range, it is taken to represent the laryngeal frequency at the time $(t^n_{pulse} + t^{n+1}_{pulse})/2$; otherwise, the duration between the pulses is considered to correspond to an unvoiced region of speech. The Lx limits are $\geq$50Hz for male speakers, and $\geq$120Hz for female speakers. There must be at least three laryngograph pulses in each voiced section. This final restriction is imposed to remove the few errors when a 'pulse' in the laryngograph data is formed by events other than glottal activity.

The accuracy with which Lx can be determined by this method is limited by the time quantisation in sampling the laryngograph signal. Each value of Lx has an error of $F_s/(F_s^2/Lx^2 - 1)$ Hz, where F_s is the sampling frequency.

2.2 Comparative Evaluation

A database containing approximately 5 minutes of speech was used for the evaluation. It was formed from sentences read by one male and one female, and was biased towards utterances containing voiced fricatives, nasals, liquids and glides, since PDAs generally find these difficult to analyse. The quantisation error in determining Lx has a mean of 0.80Hz and population standard deviation of 0.34Hz for the male speaker, and a mean of 3.33Hz and standard deviation of 0.86Hz for the female speaker. This error cannot be compensated for and affects the evaluation results for the various PDAs shown in Table 1. Durations of unvoiced or silent regions incorrectly classified as voiced by a PDA, and durations of voiced sections erroneously classified as unvoiced, are accumulated over all the utterances in the database for each speaker, and expressed as a percentage of the total duration of unvoiced (or silent) speech and voiced speech respectively. The total number of comparisons for which the difference between $F0$ and the reference Lx is higher or lower than 20% of Lx (gross errors) are expressed as a percentage of the total number of comparisons for which $F0$ and Lx represent voiced speech. The population standard deviation (p.s.d.) and the mean, absolute deviation of the Lx and $F0$ contours are given for when both represent voiced speech, and the PDA has not made a gross error.

PDA	Unvoiced in error (%)	Voiced in error (%)	Gross errors High (%)	Gross errors Low (%)	Absolute deviation (Hz) mean	Absolute deviation (Hz) p.s.d.
CPD	18.11	19.89	4.09	0.64	2.94	3.60
FBPT	3.73	13.90	1.27	0.64	1.86	2.89
HPS	14.11	7.07	5.34	28.15	3.25	3.21
IPTA	9.78	17.45	1.40	0.83	2.67	3.37
PP	7.69	15.82	0.22	1.74	2.64	3.01
SRPD	4.05	15.78	0.62	2.01	1.78	2.46
eSRPD	4.63	12.07	0.90	0.56	1.40	1.74
CPD	31.53	22.22	0.61	3.97	6.39	7.61
FBPT	3.61	12.16	0.60	3.55	5.40	7.03
HPS	19.10	21.06	0.46	1.61	4.59	5.31
IPTA	5.70	15.93	0.53	3.12	4.38	5.35
PP	6.15	13.01	0.26	3.20	6.11	6.45
SRPD	2.35	12.16	0.39	5.56	4.14	5.51
eSRPD	2.73	9.13	0.43	0.23	4.17	5.13

Table 1: PDA evaluation for male speech (top) and female speech (bottom)

Any PDA producing an $F0$ contour suitable for intonation analysis must perform consistently between male and female speech. The resultant contour must accurately determine voicing so that pitch accents are not left undetected and gross $F0$ errors must be minimal for subsequent processing. CPD and HPS are therefore unsuitable algorithms for the task in hand. Of those remaining, FBPT and SRPD form the best voiced/unvoiced classifications. SRPD would be the most suitable algorithm if the voiced/unvoiced classification performance could be improved and the number of gross ($F0$ too low) errors reduced.

3 ENHANCED SUPER RESOLUTION PITCH DETERMINATOR (*e*SRPD)

The speech is initially low-pass filtered. Frames of data for which an $F0$ estimate is required at the time of sample $s(1)$ are divided into three consecutive segments of n samples,

$$\begin{aligned} x_n &= \{x(i) = s(i-n) \mid i = 1 \text{ to } n\} \\ y_n &= \{y(i) = s(i) \mid i = 1 \text{ to } n\} \\ z_n &= \{z(i) = s(i+n) \mid i = 1 \text{ to } n\} \end{aligned} \quad (1)$$

Frames at the beginning of an utterance, for which x_n is not fully defined, are classified as 'silent'; likewise frames

at the end of an utterance, for which y_n and z_n are not fully defined.

The value of n is optimised so that each segment occupies a fundamental period. The optimisation selects a value of n within a limited range, N_{min} to N_{max} samples, which is directly related to the expected range of $F0$ values for a given speaker. The minimum and maximum values of the sample sets $x_{N_{min}}$ and $y_{N_{min}}$ are determined. If the sum of these (absolute) values is less than some preset threshold for either set, then the frame is classified as 'silent'. Otherwise, the coefficient $p_{x,y}(n)$ is determined for the values of n within the limited range in steps of a decimation factor L ($L = 4$ in this investigation).

$$p_{x,y}(n) = \frac{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL).y(jL)}{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL)^2 . \sum_{j=1}^{\lfloor n/L \rfloor} y(jL)^2} \qquad (2)$$

where $\{n = N_{min} + i.L \mid i = 0, 1, \ldots; N_{min} \leq n \leq N_{max}\}$

$p_{x,y}(n)$ is invalid if the number of zero-crossings in $x_n + y_n$ is less than 4. The locations of local maxima in $p_{x,y}(n)$ with values above an adaptive threshold (as described by Medan *et al.* [3]) form candidates for the optimum value of n. If no candidates for the fundamental period are found, the frame is classified as 'unvoiced'. Otherwise, the frame consists of 'voiced' speech, and a second coefficient, $p_{y,s}(n)$ is determined for all the fundamental period candidates. Those candidates for which $p_{y,s}(n)$ also exceeds the threshold value are given a score of 2, while the others are given a score of only 1. Candidates with a higher score are more likely to represent the true fundamental period. If there are one or more candidates with a score of 2, then all those with a score of only 1 are removed from the list of candidates and ignored. Following this, if there is only one candidate (with a score of either 1 or 2,) the candidate is assumed to be the best estimate of the fundamental period for that frame. Otherwise, the candidates are listed in order of increasing fundamental period. The candidate at the end of this list is selected to represent a fundamental period of n_M, and the m'th candidate a period n_m. Another coefficient, $q(n_m)$, is calculated for each candidate, where $q(n_m)$ is the correlation coefficient between two segments of length n_M spaced n_m apart.

$$q(n_m) = \frac{\sum_{j=1}^{n_M} s(j - n_M).s(j + n_m)}{\sum_{j=1}^{n_M} s(j - n_M)^2 . \sum_{j=1}^{n_M} s(j + n_m)^2} \qquad (3)$$

The first coefficient $q(n_1)$ is then assumed to be the ideal. If a subsequent $q(n_m)$ exceeds this ideal when multiplied by 0.77 (arbitrary) then it is in turn assumed to be the new ideal. The candidate for which the value of $q(n_m)$ is believed to be ideal is taken as the best estimate for the fundamental period of the frame being analysed.

In the case when there is only one fundamental period candidate with a score of 1 and no candidates with a score of 2, there is only a small probability that the candidate correctly represents the true fundamental period of the frame. If, in such cases, the previous frame was classified as either 'silent' or 'unvoiced', then the $F0$ value describing the current, 'voiced' frame is held until the state of the subsequent frame is known. If this next frame is also not classified as 'voiced', then the frame whose $F0$ value is on hold is an isolated frame which is highly unlikely to be voiced. It is therefore re-classified as 'unvoiced'. Otherwise, the held $F0$ value is assumed to be a sufficiently good $F0$ estimate for that frame.

Biasing is applied to the coefficients $p_{x,y}(n)$ and $p_{y,s}(n)$ for values of n where the fundamental period of a new frame is expected to lie, if the two previously analysed frames were classified as 'voiced,' if the $F0$ value of the previous frame is not being temporarily held, and if the fundamental frequency of the previous frame f_0^{r-1} is less than $\frac{7}{4}$ times the fundamental frequency of its preceding voiced frame f_0^{r-2}, and greater than $\frac{5}{8} f_0^{r-2}$, ie. if it is highly probable that the fundamental period estimate of the previous frame is not erroneous. The fundamental period of the new frame n_0^r is expected to lie within the range of n closest to n_0^{r-1} for which the set of $p_{x,y}(n)$ from the previous frame are greater than zero (Fig. 2). The coefficients $p_{x,y}(n)$ and $p_{y,s}(n)$ are doubled for values of n in this range. This effectively applies a bias on the location of a maxima in the region of the fundamental period for the previous frame to form a candidate for the fundamental period of the current frame. Note, however, that the voiced/unvoiced decision is based on the presence or absence of local maxima in $p_{x,y}(n)$ which exceed the adaptive threshold. The biasing will therefore tend to increase the percentage of unvoiced regions being incorrectly classified as 'voiced'. In order to minimise this undesirable side effect, if the unbiased coefficient $p_{x,y}(n)$ does not exceed the threshold for the candidate believed to the best estimate of the frame fundamental period, then the $F0$ value for that frame is held until the state of the subsequent frame is known. If this next frame is classified as 'silent' or 'unvoiced', the former frame is re-classified as 'unvoiced'.

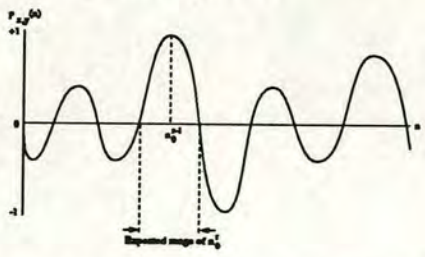


Fig. 2: Example Set of $p_{x,y}(n)$ from Previous Frame

The remainder of the processing to refine the accuracy of the $F0$ estimate is as described by Medan *et al.* [3]. The results shown in Table 1 for the enhanced SRPD algorithm (eSRPD) demonstrate the effects of these modifications on

the performance of the algorithm.

4 POST-PROCESSING OF F0

The number of short sections of gross $F0$ errors that occur during $F0$ extraction are reduced by applying a non-linear smoother [7] spanning 17 frames (with a 6.4ms interval between frames). Linear smoothing, using a hanning window, is also applied to remove small perturbations in the contour, and reduce the effect of any remaining quantisation errors. A window of length l takes l consecutive values and weights the n'th value by a factor $h(n)$. The output of the linear smoother is the sum of the weighted values.

$$h(n) = \frac{1}{l+1}\left(1 - \cos\frac{2\pi n}{l+1}\right) \text{ for } 1 \leq n \leq l \qquad (4)$$

A smoothed $F0$ contour may exhibit an overall downward trend in $F0$ during the course of an utterance. Such declination may result in two accents which have the same perceived pitch having different $F0$ values. Compensation for this effect is attempted by initially applying least median of squares regression [10] to all the local maxima and to all the local minima in the $F0$ contour. An average line is taken between the two resultant linear models and used as an estimate of the declination. Declination-compensation is only applied if this average line has a negative slope. The mean and the population standard deviation of the pre-declination-compensated $F0$ contour are retained by using frequency shifting and scaling.

Z-score normalisation [9] is applied to the declination-compensated contour to enable different $F0$ values from different speakers to correspond to the same phonological pitch. An observed $F0$ value is expressed as a multiple of a measure of dispersion relative to the mean $F0$. The normalised $F0$ value is given by,

$$F0_{norm} = \frac{F0_{input} - \overline{F0}}{\sigma} \qquad (5)$$

where $\overline{F0}$ is the long-term mean $F0$ for a given speaker, and σ is the long-term population standard deviation.

Short breaks in continuity of the normalised contour are filled by linear interpolation. Breaks are only filled if they have a duration of less than 80ms and if the jump in $F0_{norm}$ across the break is less than $\sigma/2$. The contour is then smoothed again through a hanning window spanning 17 frames.

This series of processes forms a schematised $F0$ contour which is used in intonation analysis for foreign language teaching [8].

5 CONCLUSIONS

Laryngograph signals, which are recorded simultaneously with speech, have been used to derive 'reference' Lx contours. These have been used to evaluate a selection of pitch determination algorithms. The evaluation shows the enhanced super resolution pitch determinator (eSRPD) to offer improved performance relative to the other PDAs reported in this study, with less than 1.5% gross $F0$ errors and less than 16.7% of speech classified as voiced or unvoiced incorrectly. The new eSRPD algorithm has been shown to perform well independently of a speaker's sex and is unlikely to leave pitch accents undetected. This PDA is therefore the most suitable of those investigated for applications in computer aided intonation teaching where a raw $F0$ contour is smoothed, compensated for the declination of $F0$ over an utterance, normalised for speaker differences in long-term $F0$ level and range, interpolated over small gaps, and smoothed again, to form a schematised $F0$ contour. The schematised contour is used as one of the main inputs to an automatic system for intonation teaching.

REFERENCES

[1] B. Gold and L. Rabiner. Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustical Society of America*, 46(2, part 2):442–448, 1969.

[2] W.H. Hess. *Pitch Determination of Speech Signals: Algorithms and Devices*. Springer-Verlag, Heidelberg, Germany, 1983.

[3] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, ASSP-39(1):40–48, 1991.

[4] A.M. Noll. Cepstrum pitch determination. *Journal of the Acoustical Society of America*, 41(2):293–309, 1967.

[5] A.M. Noll. *Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate*, volume 19 of *Symposium on Computer Processing in Communication*, pages 779–797. Polytechnic Institute of Brooklyn Microwave Research Institute, New York, 1970.

[6] M.S. Phillips. A feature-based time domain pitch tracker. *Journal of the Acoustical Society of America*, 77:S9–S10(A), 1985.

[7] L.R. Rabiner, M.R. Sambur, and C.E. Schmidt. Applications of non-linear smoothing algorithms to speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-23(6):552–557, 1975.

[8] E.J. Rooney, S.M. Hiller, J. Laver, and M.A. Jack. Prosodic features for automated pronunciation improvement in the SPELL system. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 413–416, Banff, Canada, 1992.

[9] P. Rose. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication*, 6(4):343–352, 1987.

[10] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[11] M.R. Schroeder. Period histogram and product spectrum: New methods for fundamental frequency measurement. *Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

[12] B.G. Secrest and G.R. Doddington. An integrated pitch tracking algorithm for speech systems. In *Proc. IEEE ICASSP-83*, pages 1352–1355, Boston, 1983.

Speech Communication 13 (1993) 333–342
North-Holland

An investigation of acoustic events related to sentential stress and pitch accents, in English

Paul C. Bagshaw

ATR Interpreting Telephony Research Laboratories Centre of Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland, UK

Received 7 January 1993
Revised 2 June 1993

Abstract. An algorithm is described which abstracts acoustic parameters of a speech waveform to automatically transcribe sentential stress and pitch movements. The waveform acoustics used are duration, energy and fundamental frequency. The abstractions described aim to isolate the prosodically imposed variations in these parameters. A method of syllabification from acoustic parameters is presented. The prominence of each syllable is determined using the automatic process described and the resultant transcription is compared with a hand-labelled prosodic transcription. The agreement level of 61.6% suggests that acoustic parameters other than those already used by the algorithm may be available to the human labeller.

Zusammenfassung. Es wird ein Algorithmus beschrieben, der die akustischen Parameter aus einem Sprachsignal herauskristallisiert, um automatisch die Bewegungen der Satz- und Wortbetonung zu umschreiben. Die verwendeten, akustischen Parameter sind die Zeit, die Energie und die Grundfrequenz. Die beschriebenen Ableitungen dienen zur Isolierung der durch die Betonung erzwungenen Variationen dieser Paramcter. Es wird eine Methode der Silbenbildung anhand der akustischen Parameter beschrieben. Die Prominenz jeder Silbe wird anhand des beschriebenen, automatischen Verfahrens bestimmt und die sich daraus ergebende Umschreibung wird mit einer handgeschriebenen Umschreibung der Betonung verglichen. Die Übereinstimmung von 61,6% lässt vermuten, dass zur menschliche Umschreibung andere akustische Parameter verwendet werden als die im Algorithmus verwendeten.

Résumé. On décrit un algorithme qui transforme des paramètres acoustiques du signal de parole en éléments abstraits permettant de transcrire automatiquement l'accent de phrase et les mouvements intonatifs. Les paramètres acoustiques utilisés sont la durée, l'énergie et la fréquence fondamentale. Les traits abstraits qui en sont dérivés visent à isoler, au sein de ces paramètres, les variations prosodiquement imposées. On présente une méthode de syllabification à partir des paramètres acoustiques. La prominence de chaque syllabe est déterminée de façon automatique et la transcription résultante est comparée à une transcription manuelle. Le taux de concordance de 61.6% suggère que l'étiqueteur humain utilise probablement d'autres paramètres que ceux pris en compte par l'algorithme.

Keywords. Prosodic labelling; syllabification; stress; pitch movements.

1. Introduction

We wish to automatically label prosodic events in English speech. Prosodic events annotated by phoneticians vary considerably from labeller to labeller and may be annotated inconsistently within any labeller's transcription. Pickering et al. (in press) show that two transcribers select the same prosodic label (level, fall, rise, fall–rise, rise–fall, stressed but unaccented or unstressed) for 72% of syllables. This paper presents a method of automatically transcribing prosodic events with the relative stress of any syllable and the relative height of pitch movements being described in scalar terms rather than as discrete levels. The method involves a series of abstractions of acoustic parameters, which aim to isolate the prosodic variations in duration, energy and fundamental frequency from the microprosodic variations.

Phones are grouped into syllable sized units

SSDI 0167-6393(93)E0069-8

which are assumed to only ever be perceived as a *single* prominent unit in continuous speech. Prosodic events are transcribed in terms of these units. The method used to produce the phone groups is described in Section 2. The prosodic content of each phone group is described by giving it a measure of its relative stress in the utterance and, if the group is accented, the type of pitch movement (level, fall, rise, fall–rise or rise–fall) and the relative height of the movement. In deducing the relative stress, the acoustic parameters, duration, energy and fundamental frequency (F_0), are abstracted to isolate prosodic variations. The formation of a piece-wise F_0 contour to remove microprosodic variations is described in Section 3. The piece-wise sections crossing each syllable are abstracted into one of five types of pitch movement. Each syllable is then labelled as either prominent (sententially stressed) or not prominent (sententially unstressed), and if it is found to be prominent and pitch salient, it is labelled as accented (Section 4). These labels are compared with those transcribed by hand.

2. Syllabification from acoustic parameters

The algorithm described below is used to group phones into syllable sized units using the phone boundary and label information given by an automatic segmentation system, and a sonorant energy contour. The segmentation data gives the location of phones and classifies those phones which are potential syllable nuclei (vowels and syllabic consonants). The sonorant energy contour of the utterance is used to determine their grouping. The syllabification of speech from acoustic parameters groups phones according to the manner in which the speaker formed the utterance rather than that dictated by a set of abstract phonological rules.

2.1. Syllabification algorithm

The energy contour for a speech waveform (sampled at 20 kHz) is calculated from 20 ms frames at 5 ms intervals so that values are synchronised with the cepstral coefficients and lower three formant frequencies used in the auto-segmentation process. Each frame is passed through a Blackman–Harris window (Harris, 1978) and the frequency bins of an amplitude spectrum (512-point FFT) corresponding to the range 50 Hz–2 kHz are accumulated. These energy values are expressed in decibels with respect to the maximum frame energy in the utterance to form an utterance-normalised sonorant energy contour. The contour is processed by a three-frame median filter and five-frame Hanning window smoother (Rabiner et al., 1975) in order to remove small perturbations which arise during frames of speech with low fundamental frequency (typically less than two pitch periods per analysis frame).

All minima in the energy contour are located and form candidates for syllable boundaries. The areas of silence identified by the auto-segmentation are respected and the minima within such areas are believed to be due to variations in background noise. Each boundary between a silence and a phone label is taken as either the beginning or the end of a syllable. The nearest candidate (energy minimum) to such a boundary is therefore moved to align with the boundary, and all those residing within the area of silence are disregarded. The regions between all the remaining energy minima are taken to be potential syllables, with a start time given by the location of the nearest minimum on the left-hand side, and the location of the nearest minimum on the right-hand side giving the stop time. We then determine whether or not the location of each of these potential syllables overlaps more than 50% of any auto-segmented vowel. If a potential syllable overlaps more than one vowel segment in this way, then the vowel segment with the maximum sonorant energy is taken to be the nucleus of the syllable. If no such overlap occurs, then we determine whether or not the location of the potential syllable overlaps more than 50% of one of the possible syllabic consonant segments/l, m, n/ (/r/ is included for American English). Again, if the potential syllable overlaps more than one of these, the one with the maximum sonorant energy is selected as the syllabic nucleus. If there is insufficient overlap, the region between the minima does not correspond to a syllable unit and

either the left-hand side or the right-hand side minimum is disregarded as a syllable boundary candidate – whichever has the highest energy and does not correspond to a phone/silence boundary. The newly formed region is then taken to be a potential syllable and the process is repeated. The resultant syllabification has boundaries located at positions of minimum sonorant energy in the utterance. These boundaries may be aligned with the auto-segmentation by moving each syllable boundary to the nearest phone boundary.

An example of the syllabification produced using this method is shown in Figure 1 for the word "International", in the phrase "... at the Kyoto International Conference Centre". The upper part shows the speech waveform and its corresponding phonemic transcription (obtained by auto-segmentation) using MRPA labels (a machine-readable phonemic alphabet). Phone boundaries are shown by dotted lines and the continuous lines show the syllable boundaries derived on a phonological basis. The lower part gives the utterance-normalised sonorant energy contour and transcription-aligned syllable boundaries derived using the algorithm described above. The phonologically based syllabification gives five syllables, /ɪn–tɚ–næ–ʃə–nəl/, but the acoustically based syllabification gives only four syllables, /ɪn–tɚ–næʃ–ənəl/. The final syllable, /–ənəl/, may initially appear to illustrate an error in the syllabification algorithm (that of a *missing* syllable boundary). However, the sonorant energy contour shows only one peak of intensity in this section of speech, and it is transcribed by a phonetician as /–nəl/. There is an error in the automatic segmentation of the speech, but not in its subsequent syllabification.

2.2. *Evaluation procedure*

A database of 453 utterances from the English language ATR conference-registration dialogues

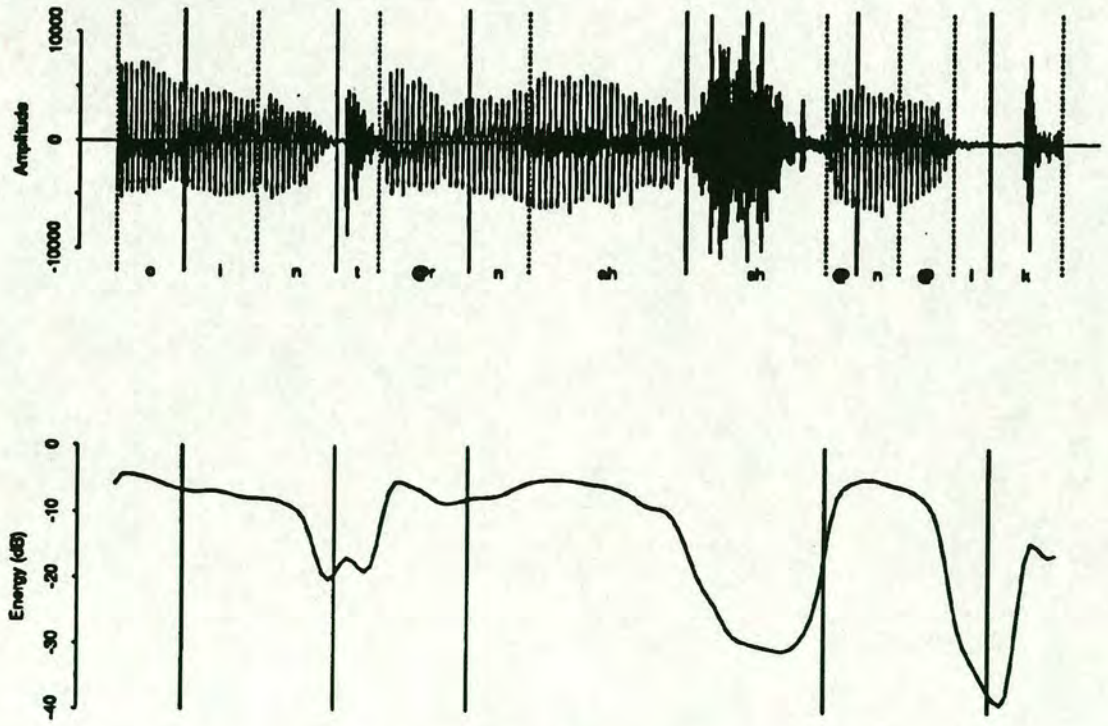


Fig. 1. Syllabification of "International".

(see (Campbell, 1992) for a brief description of this material) has been syllabified using the above algorithm and by hand using a phonologically based syllabification (Bagshaw and Williams, 1992).

In order to compare the human and machine syllabifications, it is necessary to define when they are regarded as "sufficiently similar" and when they are not. At the most stringent level of comparison, a syllable located automatically is regarded as correct only if its boundaries (both at the beginning and the end of the syllable) match exactly with those defined on a phonological basis. This method of comparison is unrepresentative of the algorithm's performance as just one miss-matched boundary would correspond to two miss-matched syllables. Furthermore, the boundaries between syllables vary even between phonological definitions – only, the syllable nuclei are well defined. Therefore, the following procedure is used to evaluate the automatic syllabification algorithm.

Initially assume that all of the syllables located automatically are *extra* syllables – i.e., that each auto-syllable is not the single match of a phonologically defined syllable. Also assume that all of the phonologically defined syllables are *missing* in the automatic syllabification – i.e., that each hand-syllable is not matched by any syllables located automatically. The phonetic transcription of every syllable is known. By definition, there will be at least one vowel and/or syllabic consonant in each syllable (defined either automatically or on a phonological basis). However, the phonetic units which form each syllable nucleus are not stipulated. The comparison proceeds by considering each hand-syllable in turn. Locate the first vowel or syllabic consonant in the hand-syllable. Then consider each auto-syllable from the beginning of the utterance to determine whether or not it overlaps with the potential nucleus. If the auto-syllable does overlap and if it has previously been assumed to be an *extra* syllable, then mark it as not being an *extra* syllable, mark the hand-syllable as *not missing*, and consider the next hand-syllable in the utterance. Otherwise, repeat the search for a matching auto-syllable using another potential nucleus in the hand-syllable until all possible nuclei have been considered.

There is a large level of agreement between the two resultant syllabic domains (see Table 1). The *missing* syllable boundaries are due to the occurrences of vowel/vowel boundaries between which there is no valley in the sonorant energy. When this case arises, often one of the vowels is a schwa; for example, the phonological syllabification of "my address" as /m aɪ–ə–d r ɛ s/ can be grouped on an acoustic basis as /m aɪ ə–d r ɛ s/. Conversely, *extra* syllable boundaries occur when the sonorant energy dips within the phonologically based syllable at a vowel/vowel boundary or vowel/syllabic consonant boundary; for example the phones in "tour" /tuɚ/ can be grouped as /tu–ɚ/ on an acoustic basis, and for the word "forms" /fɔrmz/ phones are grouped as /fɔ–rmz/ when its pronunciation tends towards that of "forums" with schwa deletion, but a fall in energy remaining between /ɔ/ and /m/.

The above algorithm forms groups of phones with a vowel or syllabic consonant as the nucleus of each. There are no vowel/vowel pairs or vowel/syllabic consonant pairs with dips in sonorant energy between them, in any of the groups. The boundaries between groups are positioned at the point of minimum sonorant energy between nuclei (aligned to the nearest phone boundary). It is assumed that each such grouping of phones can only ever be perceived as a single prominent unit in continuous speech. These units have been

Table 1
Comparison of phonologically based and acoustically based syllabifications

Total number of syllables		Match	Missing	Extra
from a phonological basis	from the acoustics			
7299 (100.0%)	7011 (96.1%)	6980 (95.6%)	−319 (4.4%)	+31 (0.4%)

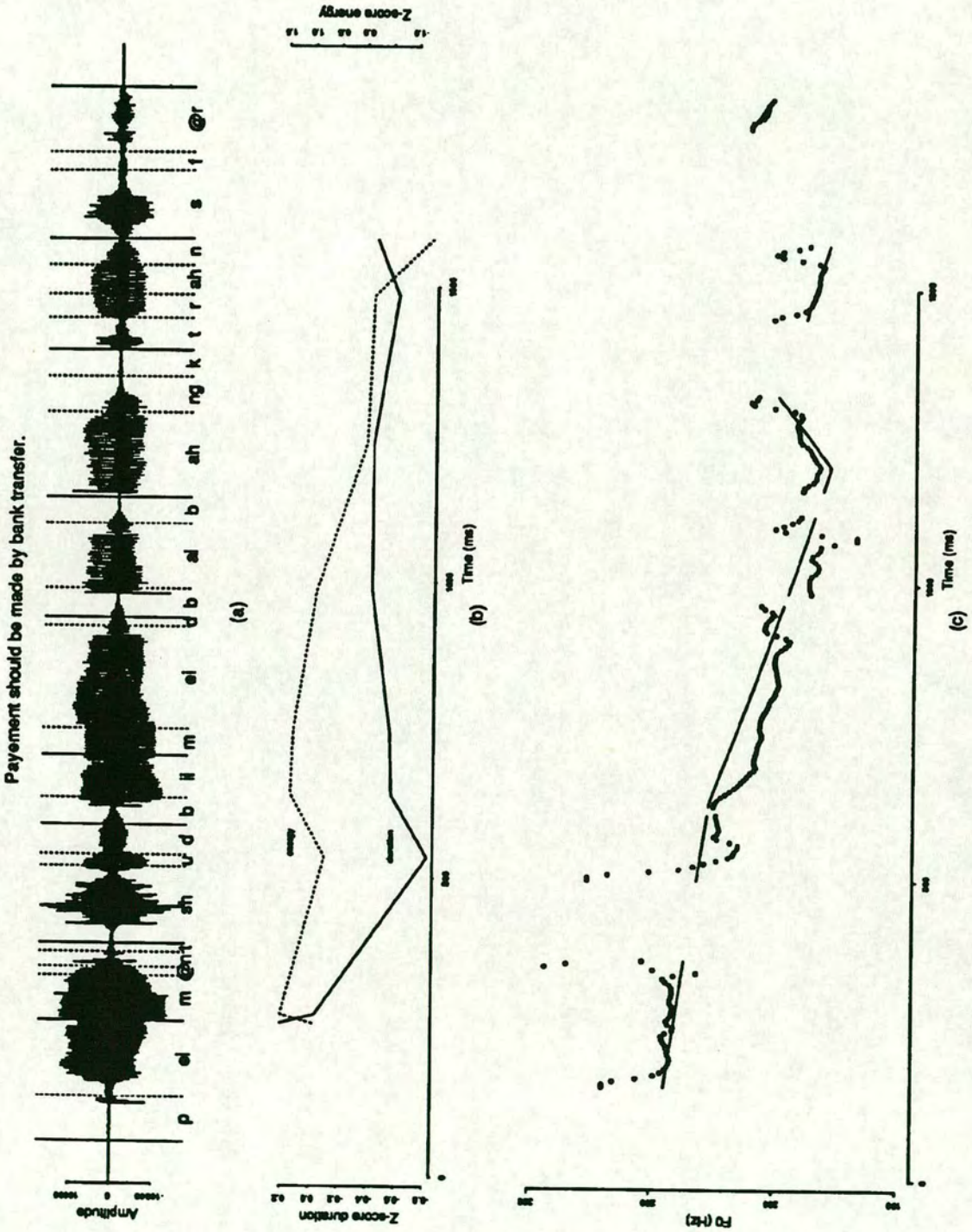


Fig. 2. Example of the abstraction of acoustic features related to prosodic events.

shown to correlate closely with syllables defined using phonologically based rules. They are, therefore, referred to as "syllables."

2.3. Duration and energy measures

Duration and sonorant energy measures are used in determining the prominence of each syllable. The duration and energy variations are mainly attributed to phone type. These parameters are therefore Z-score normalised with respect to the phone type in order to compensate for segmental variations (Campbell, 1990). The mean duration and energy and their population standard deviations are determined for each phone type from a training database of 200 utterances which provide almost total coverage of all permissible demi-syllables in English (Laver et al., 1988). The training database contains speech spoken by the same speaker in the ATR database of conference-registration dialogues and is phonetically transcribed by hand. This data is also used to train the Hidden Markov models for the automatic segmentation system. The Z-score normalisation of a phone's duration and sonorant energy simply involves subtracting the mean value and dividing by the population standard deviation for that phone type. The maximum phone-normalised duration and the maximum phone-normalised sonorant energy within a syllable are used in determining its relative prominence. (Campbell (1990) uses a similar normalisation on a phone by phone basis.)

An example of these processes is shown in Figure 2. Part (a) shows the speech waveform and its corresponding phonemic transcription (obtained by auto-segmentation). Phone boundaries are shown by dotted lines and the continuous lines show the syllable boundaries derived using the syllabification algorithm in Section 2.1. The maximum phone-normalised duration and energy for each syllable are given in part (b). These will be discussed further in Section 4.

3. The formation of a piece-wise F_0 contour

A fundamental frequency (F_0) contour produced by a pitch determination algorithm (PDA) can be expected to contain values which are inaccurate, such as instances of pitch octave errors. Any PDA will also make erroneous classifications of sections of speech as voiced or unvoiced. An enhanced version of the PDA described by Medan et al. (1991) (which is used in this study) has been found to estimate F_0 with consistently less than 1% gross pitch errors and less than 16% of speech classified as voiced or unvoiced incorrectly, when compared with laryngeal frequency estimates, F_x (Bagshaw et al., 1993). In order to eliminate the majority of octave errors and reduce micro-perturbations, the contour is initially processed by a three-frame median filter and three-frame hanning window smoother (Rabiner et al., 1975). The frames of speech analysed by the PDA are in synchronisation with those used in calculating the sonorant energy contour and in the auto-segmentation. The resultant contour is an excellent estimate of the fundamental frequency of the speech waveform, but it does not form a descriptor of utterance intonation alone as microprosodic variations are also present. A process of piece-wise linear stylisation of the contour aims to eliminate such variations.

The algorithm used to perform the stylisation is based on the technique described by Scheffers (1988) and incorporates the robust least median of squared residuals regression (LMedS) (Rousseeuw and Leroy, 1987). The F_0 values describing the contour (excluding values which equal zero to represent unvoiced speech) are converted to the semitone scale using the relationship $F_{0\,\text{semitone}} = 12 \log_2(F_{0\,\text{hertz}}/55)$. An account is given below which describes how "significant" turning-points in the F_0 contour are located, then modified to prevent contour discontinuities other than at the boundaries between unvoiced and voiced speech, and how a new, stylised contour is generated by interpolating between the turning-points.

The following process is used to identify the turning-points. Starting with the first voiced frame, LMedS analysis is applied to a window of w frames corresponding to voiced speech, where w is initially set to 5. The final frame in this window is taken to be a turning-point candidate. The F_0 value of the subsequent frame is predicted using the coefficients of the LMedS analysis. If the absolute difference between the actual

and predicted F_0 values is less than or equal to some level of permitted variation in F_0 (1 semitone), then the candidate is not a turning-point, the window length w is incremented to include the next voiced frame, and the above process is repeated. The repetition of this process terminates when the turning-point candidate is the final voiced frame in the F_0 contour. Otherwise, when the absolute difference is greater than the permitted F_0 variation, either this subsequent F_0 value constitutes some type of irregularity in the F_0 contour or the candidate could be a true turning-point. To determine which is the case, the F_0 value of the next voiced frame is also predicted. If the absolute difference between the predicted and actual values is once again greater than the permitted F_0 variation, and this situation arises for all following frames up to either the final voiced frame in the contour or such that the duration of this discontinuity (including any intermediate unvoiced frames) is greater than some minimum permitted level (100 ms), which ever occurs first, then the candidate is said to be a true turning-point. Otherwise, the length of the window w is increased to include the first frame for which the absolute difference in actual and predicted F_0 values was less than or equal to the permitted variation, but not those for which it was greater, and the LMedS analysis process is repeated. If the candidate was found to be a turning-point and if it corresponds to a voiced frame immediately preceding a frame of unvoiced speech, then the first frame of the next voiced region is also designated as a turning-point. This process is then repeated with the length of the window w reset to 5 and the first frame of the window is set to the frame of the most recent turning-point found. The first and final voiced frames of the non-stylised contour are also assigned as turning-points.

In order to ensure that discontinuities in the stylised F_0 contour only occur at unvoiced sections of speech, the fundamental frequency at each turning-point of the new contour is determined in a way which depends upon the voicing state of the frames adjacent to it. However, a discontinuity in the stylised contour is allowed within a voiced section of speech if the turning-point is an *outlier* for either of the piece-wise sections it joins. (The detection of outliers is a part of the LMedS analysis.) In such situations, the turning point is treated as being adjacent to an unvoiced frame. For any given turning-point (tp) at frame f_{tp} with original fundamental frequency $F_{0\,tp}$, the LMedS coefficients s_{tp} (slope) and i_{tp} (intercept) of the windowed points preceding the turning-point are known. The modified fundamental frequency $F'_{0\,tp}$ is given as

$$F'_{0\,tp} = \begin{cases} 0.5(s_{tp} f_{tp} + i_{tp} + s_{tp+1} f_{tp} + i_{tp+1}) \\ \quad \text{if frames } f_{tp}-1 \text{ and } f_{tp}+1 \text{ are voiced,} \\ s_{tp+1} f_{tp} + i_{tp+1} \\ \quad \text{if frame } f_{tp}-1 \text{ is unvoiced and frame } f_{tp}+1 \\ \quad \text{is voiced,} \\ s_{tp} f_{tp} + i_{tp} \\ \quad \text{if frame } f_{tp}-1 \text{ is voiced and frame } f_{tp}+1 \\ \quad \text{is unvoiced,} \\ F_{0tp} \\ \quad \text{if frames } f_{tp}-1 \text{ and } f_{tp}+1 \text{ are unvoiced.} \end{cases} \quad (1)$$

The new stylised contour is then created by linear interpolation of F_0 between each turning-point (f_{tp}, $F'_{0\,tp}$) and by resetting each frame that is unvoiced in the non-stylised contour to an unvoiced state in the new one. The resultant data is then coverted back to a Hertz scale.

An example of this piece-wise stylisation is shown in Figure 2(c). The microprosodic variations in F_0 at boundaries of voiced and unvoiced segments are eliminated whilst retaining the overall melody of the utterance.

The F_0 contours produced for the database of 453 utterances have been stylised using this method. A visual inspection has been made to compare the original F_0 contours with the piece-wise F_0 contours. The database contains 1818 syllables transcribed by hand as accented and 7316 examples of microprosodic perturbations in F_0 at voiced/unvoiced segment boundaries. The piece-wise stylisation erroneously eliminates 356 of the variations in pitch associated with an accent syllable as occurrences of short microprosodic perturbations. 610 of the microprosodic perturbations observed remain in the F_0 contour after stylisation. However, only 72 are later confused as being associated with an accented syllable by the subsequent prosodic abstraction described in Section 4.

4. Prosodic abstraction

Each piece-wise section in the stylised F_0 contour forms a possible pitch movement or part of a movement. Some piece-wise sections do not correspond to part of any pitch movement, such as those which are a direct consequence of erroneous F_0 extraction or stylisation. Thus, only those piece-wise sections which, at some time, run through any part of a syllable nuclear phone (where F_0 estimation is expected to be reliable) are treated as being part of a pitch movement. The piece-wise sections may therefore extend beyond the syllable nucleus but only those crossing the nucleus, in part or in whole, are selected. This approach compromises between using information about the movement of F_0 through vowels alone (which may be limiting for short nuclei), and using the F_0 contour of an entire syllable (where F_0 discontinuity errors may occur).

The absolute F_0 range in an utterance will vary from speaker to speaker and from utterance to utterance. F_0 piece-wise sections are, therefore, normalised for each utterance to give relative F_0 heights. The relative height of each piece-wise section is calculated by first locating a regression line which best fits the contour turning-points using LMedS analysis. A by-product of the LMedS is the standard deviation, σ_{LMedS} of the points from the resultant linear model. The absolute F_0 at each turning-point is then converted by subtracting its modelled value and dividing by the standard deviation, σ_{LMedS}. This effectively compensates for any long term declinative tendency that may be exhibited in the fundamental frequency contour, and expresses the F_0 values relative to an utterance dependent datum.

Once the relative height of each piece-wise section has been established, they are combined to form pitch movement descriptors. The pitch movement descriptors used are level, fall, rise, fall–rise and rise–fall {–,\,/,V,Λ} (Crystal, 1969). Each piece-wise section crossing any part of a syllable nucleus is classified as either level, fall or rise. Let $F_{0\text{start}}$ be the relative F_0 height at the start of the piece-wise section and that at the end of the section be $F_{0\text{end}}$. The piece-wise section is classified on the following basis:

$$\text{pitch movement} = \begin{cases} \backslash & \text{if } F_{0\text{start}} - F_{0\text{end}} > 0.75\sigma_{\text{LMedS}}, \\ / & \text{if } F_{0\text{start}} - F_{0\text{end}} < -0.75\sigma_{\text{LMedS}}, \\ - & \text{otherwise.} \end{cases} \quad (2)$$

When more than one piece-wise section crosses any particular nucleus, they are combined by initially taking all adjacent sections with the same pitch movement classification and joining them into one. A join is made by setting $F_{0\text{start}}$ to that of the first section, $F_{0\text{end}}$ to that of the second section, and reclassifying using equation (2). In the database of 453 utterances, consisting of 7299 syllables, there were only 4 syllables for which more than two sections remained after this process. These all contained some error which originated in the F_0 estimation. If there are two remaining sections (their classifications must differ), and if either is classified as level {–}, then they too are joined in the same way. Otherwise, one is a fall {\} and the other is a rise {/}. These are combined to give a single movement classified as either a fall–rise {V} or rise–fall {Λ} depending on their order, and the relative level at their mid-point is kept. Thus, for the fall–rise and rise–fall classifications, the relative height of both the onset and coda of the movement are known.

Having established the shape of the pitch movement over each syllable in this way, and with knowledge of the Z-score normalised syllable nucleus duration and energy measures, we determine if any given syllable is prominent in the utterance (sententially stressed {s}) and if it is pitch salient (accented {a}). A syllable is labelled as prominent if both its normalised duration and its normalised energy are greater than that of both its nearest neighbours (end-points being inherently lower), and if both are greater than 0.75 standard deviations from the mean value. (The value 0.75 is arbitrary.) It is similarly labelled if either the normalised duration or normalised energy measure is the maximum for the utterance. A syllable is labelled as accented using a decision filter three pitch movements wide developed by Hieronymus (1989). This categorisation of the

scalar measures of prominence based on duration and energy, to a single discrete level (stressed or unstressed) is necessary only in order to compare the automated annotation with that made by hand.

The database of 453 utterances has been automatically prosodically labelled in this way and compared with those transcribed by hand (see Table 2). The transcriptions are equal for 61.6% of the syllables. Note, however, that the balance between syllables being transcribed automatically as prominent (accented or otherwise) or as not prominent, is dependent upon an arbitrary threshold. It is therefore only possible to observe general trends from the confusion matrix. Of the unstressed (u) hand labels transcribed as either accented or stressed automatically, 216 were syllables with a schwa nucleus. This suggests that the hand transcriber may be annotating syllables as sententially stressed only if they can be lexically stressed, and that the auto-segmentation process may have more confusions in classifying a schwa than for some other vowel. There is also a noticeably large number of syllables labelled by hand as accented or stressed that are labelled automatically as unstressed, suggesting that the hand labeller may be using acoustic parameters other than those described previously in this paper. For example, syllables whose nucleus is "fully articulated" are often labelled as stressed by hand. Such measures are currently unavailable to the automatic prosodic transcription algorithm. These errors may also be accounted for by the need to refine the Hieronymus decision filter used in locating the accented syllables and the need to enhance the F_0 piece-wise stylisation procedure. 428 of the confusions between accented and non-accented syllables arise because of erroneous F_0 stylisation.

5. Conclusions

An algorithm based on the abstraction of acoustic parameters to annotate sentential stress and pitch accents has been presented. A method of grouping phones into syllables which can consist of only one prominence has been described. This forms a domain in which prosodic events are transcribed. There is a large level of agreement (95.6%) between this domain and a phonologically based syllabification. The piece-wise stylisation of a fundamental frequency contour aimed at eliminating microprosodic variations has been illustrated. Piece-wise sections of an F_0 contour are abstracted to describe pitch movements over each syllable. The extend of these movements are known relative to other pitch movements in the utterance and they are used in locating accented syllables. The prominence of each syllable has been determined using the automatic process described and have been compared with a hand-labelled prosodic transcription. An agreement level of 61.6% has been found. This level of agreement is lower than one would expect given the large body of research which has indicated the acoustic correlates of prominence in English to be duration, energy and fundamental frequency. It appears that the hand labels transcribe aspects of speech that are not apparent in the waveform acoustics used. Possible reasons for the differences in the hand and automated transcriptions have been discussed. It is an area of on-going research to establish what contributes to the er-

Table 2
Confusion matrix of prosodic transcription by hand and by automation

		Automatic label			
		a	s	u	Total
Hand label	a	566 (7.8%)	177 (2.4%)	1075 (14.7%)	1818 (24.9%)
	s	128 (1.8%)	71 (1.0%)	792 (10.9%)	991 (13.6%)
	u	404 (5.5%)	226 (3.1%)	3860 (52.9%)	4490 (61.5%)
	Total	1098 (15.0%)	474 (6.5%)	5727 (78.5%)	7299 (100.0%)

(a = accented, s = stressed but unaccented, u = unstressed).
Correct classification rate = 4497/7299 (61.6%).

rors and to determine ways of overcoming them by refining the algorithm.

Acknowledgments

Thanks to Nick Campbell, Jacqueline Vaissière, Jim Hieronymus and Peter Meer for their valuable assistance and suggestions. I am also grateful to Steve Isard and Mervyn Jack for their assistance in the preparation of this document.

References

P.C. Bagshaw and B.J. Williams (1992), "Criteria for labelling prosodic aspects of English speech", *Proc. Internat. Conf. on Spoken Language Processing, Banff, Canada*, Vol. 2, pp. 859–862.

P.C. Bagshaw, S.M. Hiller and M.A. Jack (1993), "Enhanced pitch tracking and the processing of F_0 contours for computer aided intonation teaching", *Proc. European Conf. on Speech Communication and Technology, Berlin*, Vol. 2, pp. 1003–1006.

W.N. Campbell (1990), "Evidence for a syllable-based model of speech timing", *Proc. Internat. Conf. on Spoken Language Processing, Kobe, Japan*, Vol. 1, pp. 9–12.

W.N. Campbell (1992), "Prosodic encoding of English speech", *Proc. Internat. Conf. Spoken Language Processing, Banff, Canada*, Vol. 1, pp. 663–666.

D. Crystal (1969), *Prosodic Systems and Intonation in English* (Cambridge Univ. Press, Cambridge, UK).

F.J. Harris (1978), "On the use of windows for harmonic analysis with the discrete Fourier transform", *Proc. IEEE*, Vol. 66, No. 1, pp. 51–83.

J.L. Hieronymus (1989), "Automatic sentential vowel stress labelling", *Proc. European Conf. on Speech Communication and Technology, Paris*, Vol. 1, pp. 226–229.

J. Laver, C. Bennett, I. Cohan, J. Dalby, D. Davies and M. McAllister (1988), ATR/CSTR speech database project, Status report, No. 1 (Centre for Speech Technology Research; Univ. of Edinburgh, UK).

Y. Medan, E. Yair and D. Chazan (1991), "Super resolution pitch determination of speech signals", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-39, No. 1, pp. 40–48.

B. Pickering, B. Williams and G. Knowles (in press), "Analysis of transcriber differences in the SEC", *Working with Speech*, ed. by P. Alderson and G. Knowles (Longman, London), Chapter 4.

L.R. Rabiner, M.R. Sambur and C.E. Schmidt (1975), "Applications of non-linear smoothing algorithms to speech processing", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, No. 6, pp. 552–557.

P.J. Rousseeuw and A.M. Leroy (1987), *Robust Regression and Outlier Detection* (Wiley, New York).

M.T.M. Scheffers (1988), "Automatic stylization of F_0-contours", in: *Proc. 7th FASE Symposium, Edinburgh*, ed. by W.A. Ainsworth and J.N. Holmes, Vol. 3, pp. 981–987.

Appendix D

MRPA and IPA Symbols

Symbols have been used in this thesis to represent sound units of speech. The symbols used in the body of the text are those of the alphabet of the International Phonetics Association (IPA). The limitations of the packages used in producing this thesis have forced the symbols of the Machine Readable Phonemic Alphabet (MRPA) to be used in some of the illustrative Figures (marked by an asterisk in the Figure caption). Table D.1 shows the corresponding IPA and MRPA symbols for sound units and gives example words in which the sounds occur in the exemplar languages English, French and Italian (adopted from Hiller *et al.*, 1990).

Symbol		Example Word		
IPA	MRPA	English	French	Italian
[p]	p	pea	par	copia[1]
[t]	t	tea	tard	grato[1]
[k]	k	key	car	chiaro[1]
[b]	b	bee	barre	buono[1]
[d]	d	dye	dos	dove[1]
[g]	g	guy	gare	gatto[1]
[ʧ]	ch	each	–	città[1]
[ʤ]	jh	edge	–	Giorgio[1]
[s]	s	sea	si	stanco[1]
[z]	z	zoom	zebre	snello
[ʃ]	sh	she	chat	uscire
[ʒ]	zh	beige	jeu	–
[ʦ]	ts	–	–	zio[1]
[ʣ]	dz	–	–	zero[1]
[f]	f	fan	foi	faccia[1]
[v]	v	van	vin	beve[1]
[θ]	th	thin	–	–
[ð]	dh	then	–	–
[h]	h	hat	–	–
[m]	m	me	mare	mangia[1]
[n]	n	knee	non	vano[1]
[ŋ]	ng	song	–	–
[ɲ]	ny	–	agneau	bisogno
[l]	l	lay	le	alba[1]
[ʎ]	ly	–	–	battaglia
[r]	r	ray	–	serata[1]
[ʁ]	R	–	roi	–
[j]	y	yes	billet	più
[ɥ]	yw	–	huit	–
[w]	w	way	oui	può

continued on next page...

[1]The long version of these consonant sounds are marked in MRPA by a *colon* — [bɪ] b: 'abbazia'.
[2]Centralised.
[3]British English.
[4]American English.

... continued from previous page

Symbol		Example Word		
IPA	MRPA	English	French	Italian
[i]	ii	bead	lit	pino
[ɪ]	i	bid	–	–
[ɨ]	|	bid[2]	–	–
[e]	ee	–	été	lesso
[ɛ]	e	bed	sec	letto
[ɛ̃]	en	–	vin	–
[ɑ]	aa	bard	gras	–
[a]	a	bad[3]	la	faccia
[ã]	an	–	grand	–
[æ]	ah	sad[4]	–	–
[ɜ]	@@	bird[3]	–	–
[ɚ]	@r	word[4]	–	–
[ə]	@	about	–	–
[ʌ]	uh	bud	–	–
[u]	uu	boot	loup	uno
[ʉ]	u"	boot[2]	–	–
[ʊ]	u	put	–	–
[y]	yy	–	sur	–
[ø]	eu	–	feu	–
[œ]	oe	–	coeur	–
[œ̃]	oen	–	commun	–
[o]	0	–	eau	torre
[õ]	on	–	pont	–
[ɔ]	oo	port	fort	tosto
[ɒ]	o	pot	–	–
[eɪ]	ei	bay	–	quei
[aɪ]	ai	buy	–	sai
[ɔɪ]	oi	boy	–	–
[əʊ]	ou	go	–	–
[aʊ]	au	bow	–	laurea
[ɪə]	i@	beer[3]	–	–
[ɛə]	e@	bare[3]	–	–
[ʊə]	u@	poor[3]	–	–
[ʔ]	?	*glottal stop*		
[']	'	*primary stress*		
[ˌ]	"	*secondary stress*		
[˳]	*	*tertiary stress*		

Table D.1: IPA/MRPA Symbols

References

ADAMS, C., & MUNRO, R.R. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterance of some native and non-native speakers of English. *Phonetica*, **35**, 125–156.

ATAL, B.S., & RABINER, L.R. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP–24**(3), 201–212.

AULL, A.M., & ZUE, V.W. (1985). Lexical stress detection and its application to large vocabulary speech recognition. *Pages 1549–1552 of: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. Florida.

BAER, T., LÖFQVIST, A., & MCGARR, N.S. (1983). Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques. *Journal of the Acoustical Society of America*, **73**(4), 1304–1308.

BAGSHAW, P.C. (1992). An investigation of acoustic events related to sentential stress and pitch accents, in English. *Pages 808–813 of: Proc. 4th. Australian International Conference on Speech Science and Technology.* Brisbane, Australia.

BAGSHAW, P.C. (1993). An investigation of acoustic events related to sentential stress and pitch accents, in English. *Speech Communication*, **13**(3–4), 333–342.

BAGSHAW, P.C., & WILLIAMS, B.J. (1992). Criteria for labelling prosodic aspects of English speech. *Pages 859–862 of: Proc. International Conference on Spoken Language Processing*, vol. 2. Banff, Canada.

BAGSHAW, P.C., HILLER, S.M., & JACK, M.A. (1993). Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. *Pages 1003–1006 of: Proc. 3rd. European Conference on Speech Communication and Technology*, vol. 2. Berlin.

BARRY, W.J., & FOURCIN, A.J. (1992). Levels of labelling. *Computer Speech and Language*, **6**, 1–14.

BARTKOVA, K., & SORIN, C. (1987). A model of segmental duration for speech synthesis in French. *Speech Communication*, **6**(3), 245–260.

BEAUGENDRE, F., D'ALESSANDRO, C., LACHERET-DUJOUR, A., & TERKEN, J. (1992). A perceptual study of French intonation. *Pages 739–742 of: Proc. International Conference on Spoken Language Processing*, vol. 1. Banff, Canada.

BECKMAN, M.E. (1986). *Stress and Non-stress Accent.* Dordrecht, Holland: Foris Publications.

BECKMAN, M.E., & PIERREHUMBERT, J.B. (1986). Intonation structure in Japanese and English. *Pages 255–309 of:* EWEN, C.J., & ANDERSON, J.M. (eds), *Phonology Yearbook*, vol. 3. Cambridge, U.K.: Cambridge University Press.

BOLINGER, B. (ed). (1972). *Intonation.* Harmondsworth: Penguin Books Ltd.

BOLINGER, D.L. (1958). A theory of pitch accent in English. *Word*, **14**, 109–149.

BRAZIL, D., COULTHARD, M., & JOHNS, C. (1980). *Discourse Intonation and Language Teaching.* London: Longman.

BRUCE, G. (1977). *Swedish word accents in sentence perspective.* Lund: Gleerup.

CAMPBELL, W.N. (1990). Evidence for a syllable-based model of speech timing. *Pages 9–12 of: Proc. International Conference on Spoken Language Processing*, vol. 1. Kobe, Japan.

CAMPBELL, W.N. (1992). Prosodic encoding of English speech. *Pages 663–666 of: Proc. International Conference on Spoken Language Processing*, vol. 1. Banff, Canada.

CAMPBELL, W.N., & ISARD, S.D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, **19**, 37–47.

CHAPALLAZ, M. (1979). *The Pronunciation of Italian: A Practical Introduction.* London: Bell & Hyman.

CHOMSKY, N., & HALLE, M. (1968). *The Sound Pattern of English.* New York: Harper & Row.

COHEN, A., & 'T HART, J. (1967). On the anatomy of intonation. *Lingua*, **19**, 177–192.

COUPER-KUHLEN, E. (1986). *An Introduction to English Prosody.* London: Edward Arnold (Publishers) Ltd.

CROWE, A.S. (1988). Generalised centroids: A new perspective on peak picking and formant estimation. *Pages 683–689 of:* AINSWORTH, W.A., & HOLMES, J.N. (eds), *Proc. 7th. FASE Symposium*, vol. 2. Edinburgh.

CROWE, A.S., & JACK, M.A. (1987). Globally optimising formant tracker using generalised centroids. *Electronics Letters*, **23**(19), 1019–1020.

CRYSTAL, D. (1969). *Prosodic Systems and Intonation in English.* Cambridge, U.K.: Cambridge University Press.

CRYSTAL, T.H., & HOUSE, A.S. (1988). Segmental durations in connected-speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, **83**(4), 1574–1585.

CUTLER, A., & NORRIS, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**(1), 113–121.

DALBY, J.M. (1986). *Phonetic Structure of Fast Speech in American English.* PhD dissertation, Indiana University Linguistics Club, Bloomington, Indiana.

DAUER, R.M. (1983). Stress-timing and syllable-timing reanalysed. *Journal of Phonetics*, **11**(1), 51–62.

DE PIJPER, J.R. (1983). *Modelling British Intonation: An analysis by resynthesis of British English intonation.* Dordrecht, Holland: Foris Publications.

DI BENEDETTO, M.-G., CARRARO, F., HILLER, S.M., & ROONEY, E.J. (1992). Vowels pronunciation assessment in the SPELL system. *Pages 417–420 of: Proc. International Conference on Spoken Language Processing*, vol. 1. Banff, Canada.

DISNER, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, **67**(1), 253–261.

DUDA, R.O., & HART, P.E. (1973). *Pattern Classification and Scene Analysis.* New York: Wiley.

FALLSIDE, F., & WOODS, W.A. (eds). (1985). *Computer Speech Processing.* London: Prentice-Hall International (UK) Ltd.

FANT, G., & KRUCKENBERG, A. (1988). Some durational correlates of Swedish prosody. *Pages 495–502 of:* AINSWORTH, W.A., & HOLMES, J.N. (eds), *Proc. 7th. FASE Symposium*, vol. 2. Edinburgh.

FRY, D.B. (1955). Duration and intensity as physics correlates of linguistic stress. *Journal of the Acoustical Society of America*, **27**(4), 765–768.

FRY, D.B. (1958). Experiments in the perception of stress. *Language and Speech*, **1**, 126–152.

FRY, D.B. (1965). The dependence of stress judgement on vowel formant structure. *Pages 306–311 of:* ZWIRNER, E., & BETHGE, W. (eds), *Proc. 5th. International Congress of Phonetic Sciences, University of Münster, 16–22 August 1964.* New York: Karger.

FUDGE, E.C. (1984). *English Word-Stress.* London: George Allen & Unwin.

GILLICK, L., & COX, S.J. (1989). Some statistical issues in the comparison of speech recognition algorithms. *Pages 532–535 of: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. Glasgow, Scotland.

GIMSON, A.C. (1970). *An Introduction to the Pronunciation of English.* Second edn. London: Edward Arnold.

GODFREY, J.J., & BRODSKY, J.M. (1986). Acoustic characteristics of emphasis. *Journal of the Acoustical Society of America*, **80**(Supplement 1), S49(A).

GOLD, B., & RABINER, L. (1969). Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustical Society of America*, **46**(2, part 2), 442–448.

GRICE, M.L. (1992). *The Intonation of Interrogation in Palermo Italian: Implications for Intonation Theory.* Ph.D. thesis, University College London.

GRICE, M.L., & BARRY, W.J. (1991). Problems of transcription and labelling in the specification of segmental and prosodic structure. *Pages 66–69 of: Proc. XII'th. International Congress of Phonetic Sciences*, vol. 5. Aix en Provence, France.

GUSSENHOVEN, C. (1984). *On the Grammar and Semantics of Sentence Accents.* Dordrecht, Holland: Foris Publications.

HALLIDAY, M.A.K. (1970). *A Course in Spoken English: Intonation.* London: Oxford University Press.

HARMER, J. (1983). *The Practice of English Language Teaching.* London: Longman.

HARRIS, F.J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, **66**(1), 51–83.

HESS, W.H. (1983). *Pitch Determination of Speech Signals: Algorithms and Devices.* Heidelberg, Germany: Springer-Verlag.

HIERONYMUS, J.L. (1989). Automatic sentential vowel stress labelling. *Pages 226–229 of: Proc. 1st. European Conference on Speech Communication and Technology*, vol. 1. Paris.

HIERONYMUS, J.L., & WILLIAMS, B.J. (1991). An investigation of the relation between perceived pitch accent and automatically-located accent in British English. *Pages 1157–1160 of: Proc. 2nd. European Conference on Speech Communication and Technology*, vol. 3. Genova, Italy.

HILLENBRAND, J., & GAYVERT, R.T. (1993). Vowel classifiaction based on fundamental frequency and formant frequencies. *Journal of Speech and Hearing Research*, **36**(4), 674–699.

HILLER, S.M. (1985). *Automatic Acoustic Analysis of Waveform Perturbations.* Ph.D. thesis, University of Edinburgh, Scotland.

HILLER, S.M., ROONEY, E.J., & LAVER, J. (1990). *SPELL project speech stimuli.* Status report 1.1. Centre for Speech Technology Research, University of Edinburgh, U.K.

HILLER, S.M., ROONEY, E.J., LAVER, J., & JACK, M.A. (1993). SPELL: An automated system for computer-aided pronunciation teaching. *Speech Communication*, **13**(3–4), 463–473.

HIRST, D., & ESPESSER, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, **15**.

INKELAS, S., & ZEC, D. (1988). Serbo-Croatian pitch accent: The interaction of tone, stress and intonation. *Language*, **64**(2), 227–248.

JAMES, M. (1985). *Classification Algorithms.* London: Collins.

KENWORTHY, J. (1987). *Teaching English Pronunciation.* London: Longman.

KINGDOM, R. (1958). *The Groundwork of English Stress.* London: Longman.

KLATT, D.H. (1979). Synthesis by rule of segmental durations in English sentences. *Pages 287–299 of:* LINDBLOM, B., & OHMAN, S.E.G. (eds), *Frontiers of Speech Communication Research.* London: Academic Press Inc.

LADD, D.R. (1983). Phonological features of intonation peaks. *Language*, **59**(4), 721–759.

LADD, D.R. (1986). Intonational phrasing: the case for recursive prosodic structure. *Pages 311–340 of:* EWEN, C.J., & ANDERSON, J.M. (eds), *Phonology Yearbook*, vol. 3. Cambridge, U.K.: Cambridge University Press.

LADD, D.R. (1992). Compound prosodic domains. *Edinburgh University Department of Linguistics Occasional Paper.* [Submitted for publication in *Language*].

LADD, D.R., & CAMPBELL, W.N. (1991). Theories of prosodic structure: Evidence from syllable duration. *Pages 290–293 of: Proc. XII'th. International Congress of Phonetic Sciences*, vol. 2. Aix en Provence, France.

LADEFOGED, P. (1982). *A Course in Phonetics.* Second edn. New York: Harcourt Brace Jovanovich, Inc.

LAMEL, L.F., KASSEL, R.H., & SENEFF, S. (1986). Speech database development: Design and analysis of the acoustic-phonetic corpus. *Pages 100–109 of:* BAUMANN, L.S. (ed), *Proc. DARPA Speech Recognition Workshop, Palo Alto, California, 19–20 February 1986.* McLean, Virginia: Science Applications International Corporation.

LAVER, J. (1980). *The Phonetic Description of Voice Quality.* Cambridge, U.K.: Cambridge University Press.

LAVER, J., BENNETT, C., COHEN, I., DALBY, J., DAVIES, D., & MCALLISTER, M. (1988). *ATR/CSTR speech database project.* Status report 1. Centre for Speech Technology Research, University of Edinburgh, U.K.

LAVER, J., ALEXANDER, M., BENNETT, C., COHEN, I., & DAVIES, D. (1989). *Speech segmentation criteria for ATR/CSTR.* Tech. rept. Centre for Speech Technology Research, University of Edinburgh, U.K.

LEA, W.A. (1974). An algorithm for locating stressed syllables in continuous speech. *Journal of the Acoustical Society of America*, **55**(2), 411(A).

LEA, W.A. (1980). Prosodic aids to speech recognition. *Chap. 8 of:* LEA, W.A. (ed), *Trends in Speech Recognition.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

LEFÈVRE, J.-P., HILLER, S.M., ROONEY, E.J., LAVER, J., & DI BENEDETTO, M.-G. (1992). Macro and micro features for automatic pronunciation improvement in the SPELL system. *Speech Communication*, **11**(1), 31–44.

LEHISTE, I. (1970). *Suprasegmentals.* Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

LEHISTE, I., & PETERSON, G.E. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, **31**(4), 428–435.

LIBERMAN, M.Y. (1975). *The Intonation System of English.* PhD dissertation, Massachusetts Institute of Technology.

LIBERMAN, M.Y., & PRINCE, A.S. (1977). On stress and linguistic rhythm. *Linquitic Inquiry*, **8**, 249–336.

LIEBERMAN, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, **32**(4), 451–454.

LIEBERMAN, P. (1965). On the acoustic basis of the perception of intonation by linguists. *Word*, **21**, 40–45.

LINDBLOM, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, **35**(11), 1773–1781.

MACCARTHY, P. (1975). *The Pronunciation of French.* London: Oxford University Press.

MADDIESON, I. (1985). Phonetic Cues to Syllabification. *In:* FROMKIN, V.A. (ed), *Phonetic Linguistics (essays in honor of P.Ladefoged).* London: Academic Press Inc.

MADSEN, H.S. (1983). *Techniques in Testing.* Oxford: Oxford University Press.

MARKEL, J.D., & GRAY, JR., A.H. (1976). *Linear prediction of speech.* Heidelberg, Germany: Springer-Verlag.

MARTINEAU, R., & MCGRIVNEY, J. (1973). *French Pronunciation.* London: Oxford University Press.

MCINNES, F.R., ARIKI, Y., & WRENCH, A.A. (1989). Enhancement and optimisation of a speech recognition front end based on hidden Markov models. *Pages 461–464 of: Proc. 1st. European Conference on Speech Communication and Technology*, vol. 2. Paris.

MCINNES, F.R., CARRARO, F., HILLER, S.M., & ROONEY, E.J. (1992). Evaluation and optimisation of a segmenter for a PC-based pronunciation teaching system. *Pages 109–116 of: Proc. Institute of Acoustics Conference on Speech and Hearing*, vol. 14, part 6. Windermere, England.

MEDAN, Y., YAIR, E., & CHAZAN, D. (1991). Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, **ASSP–39**(1), 40–48.

MERMELSTEIN, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, **58**(4), 880–883.

MERTEN, P. (1987). *L'intonation du Français: De la Description Linguistique à la Reconnaissance Automatique.* Doctorale dissertatie, Katholieke Universiteit Leuven. (in French).

MORTON, J., & JASSEM, W. (1965). Acoustic correlates of stress. *Language and Speech*, **8**(3), 159–181.

NAKATANI, L., & ASTON, C.H. (1978). Perceiving stress patterns of words in sentences. *Journal of the Acoustical Society of America*, **63**(Supplement 1), S55(A).

NOLL, A.M. (1967). Cepstrum pitch determination. *Journal of the Acoustical Society of America*, **41**(2), 293–309.

NOLL, A.M. (1970). *Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate.* Symposium on Computer Processing in Communication, vol. 19. New York: Polytechnic Institute of Brooklyn Microwave Research Institute. Pages 779–797.

O'CONNOR, J.D., & ARNOLD, G.R. (1973). *Intonation of Colloquial English.* Second edn. London: Longman.

PETERSON, G.E., & BARNEY, H.L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, **24**(2), 175–184.

PHILLIPS, M.S. (1985). A feature-based time domain pitch tracker. *Journal of the Acoustical Society of America*, **77**, S9–S10(A).

PICKERING, B., WILLIAMS, B., & KNOWLES, G. (1994). Analysis of transcriber differences in the SEC. *Chap. 4 of:* KNOWLES, G., & ALDERSON, P.R. (eds), *Working with Speech: The Computational Analysis of Formal British English Speech.* London: Longman.

PIERCE, J.R. (1962). *Symbols, Signals and Noise: The Nature and Process of Communication.* London: Hutchinson.

PIERREHUMBERT, J.B. (1980). *The Phonology and Phonetics of English Intonation.* PhD dissertation, Massachusetts Institute of Technology.

PIERREHUMBERT, J.B., & BECKMAN, M.E. (1988). *Japanese Tone Structure.* Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

PIKE, K. (1945). *The Intonation of American English.* Ann Arbor, MI: University of Michigan Press. [Excerpts reprinted in Bolinger (1972)].

PRESS, W.H., FLANNERY, B.P., TEUKOLSKY, S.A., & VETTERLING, W.T. (1988). *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge, U.K.: Cambridge University Press.

PRICE, G. (1991). *An Introduction to French Pronunciation.* Oxford, England: Blackwell.

PRICE, P.J., WIGHTMAN, C.W., OSTENDORF, M., & BEAR, J. (1990). Evidence for a syllable-based model of speech timing. *Pages 13–16 of: Proc. International Conference on Spoken Language Processing*, vol. 1. Kobe, Japan.

PRICE, P.J., OSTENDORF, M., SHATTUCK-HUFNAGEL, S., & FONG, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, **90**(6), 2956–2970.

PULGRAM, E. (1970). *Syllable, Word, Nexus, Cursus.* Den Hague: Mouton.

RABINER, L.R., SAMBUR, M.R., & SCHMIDT, C.E. (1975). Applications of non-linear smoothing algorithms to speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP–23**(6), 552–557.

RABINER, L.R., CHENG, M.J., ROSENBERG, A.E., & MCGONEGAL, C.A. (1976). A comparative performance study of several pitch detection algorithms. *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP–24**(5), 399–418.

RIVERS, W.M. (1975). *A Practical Guide to the Teaching of French.* New York: Oxford University Press.

ROONEY, E.J., HILLER, S.M., LAVER, J., & JACK, M.A. (1992). Prosodic features for automated pronunciation improvement in the SPELL system. *Pages 413–416 of: Proc. International Conference on Spoken Language Processing*, vol. 1. Banff, Canada.

ROSE, P. (1987). Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication*, **6**(4), 343–352.

ROUSSEEUW, P.J. (1984). Lease median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.

ROUSSEEUW, P.J., & LEROY, A.M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

SCHEFFERS, M.T.M. (1988). Automatic stylization of F0-contours. *Pages 981–987 of:* AINSWORTH, W.A., & HOLMES, J.N. (eds), *Proc. 7th. FASE Symposium*, vol. 3. Edinburgh.

SCHROEDER, M.R. (1968). Period histogram and product spectrum: New methods for fundamental frequency measurement. *Journal of the Acoustical Society of America*, **43**(4), 829–834.

SECREST, B.G., & DODDINGTON, G.R. (1983). An integrated pitch tracking algorithm for speech systems. *Pages 1352–1355 of: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing.* Boston.

SELKIRK, E.O. (1984). *Phonology and Syntax: The Relationship between Sound and Structure.* Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

SHANNON, C.E., & WEAVER, W. (1949). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

SHEARME, J.N., & HOLMES, J.N. (1962). An experimental study on the classification of sounds in continuous speech according to their distribution in the formant 1-formant 2 plane. *Pages 234–240 of:* SOVIJARVI, A., & AALTO, P. (eds), *Proc. 4th. International Congress of Phonetic Sciences, University of Helsinki, 4–9 September 1961.* Den Hague: Mouton.

SILVERMAN, K.E.A. (1987). *The Structure of Processing of Fundamental Frequency Contours.* Ph.D. thesis, University of Cambridge, England.

SILVERMAN, K.E.A., BECKMAN, M.E., PITRELLI, J., OSTENDORF, M., WIGHTMAN, C.W., PRICE, P.J., PIERREHUMBERT, J.B., & HIRSCHBERG, J. (1992). TOBI: A standard for labelling English prosody. *Pages 867–870 of: Proc. International Conference on Spoken Language Processing*, vol. 2. Banff, Canada.

SYRDAL, A.K., & GOPAL, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**(4), 1086–1100.

'T HART, J. (1991). F0 stylization in speech: Straight lines versus parabolas. *Journal of the Acoustical Society of America*, **90**(6), 3368–3370.

'T HART, J., & COHEN, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, **1**, 309–327.

TAYLOR, P.A. (1993). *A Phonetic Model of English Intonation.* Ph.D. thesis, University of Edinburgh, Scotland.

TAYLOR, P.A., & ISARD, S.D. (1992). A new model of intonation for use with speech synthesis and recognition. *Pages 1287–1290 of: Proc. International Conference on Spoken Language Processing*, vol. 1. Banff, Canada.

TIFFANY, W.R. (1959). Nonrandom sources of variation in vowel quality. *Journal of Speech and Hearing Research*, **2**(4), 305–317.

TRANEL, B. (1987). *The Sounds of French: An Introduction.* Cambridge: Cambridge University Press.

TULLER, B., HARRIS, K.S., & KELSO, J.A.S. (1982). Stress and rate: Differential transformation of articulation. *Journal of the Acoustical Society of America*, **71**(6), 1534–1543.

VAISSIÈRE, J. (1983). Language-independent prosodic features. *Chap. 5, pages 53–66 of:* CUTLER, A., & LADD, D.R. (eds), *Prosody: Models and Measurements.* Heidelberg, Germany: Springer-Verlag.

VAISSIÈRE, J. (1988). The use of prosodic parameters in automatic speech recognition. *Pages 71–99 of:* NIEMANN, H., LANG, M., & SAGERER, G. (eds), *Recent Advances in Speech Understanding and Dialog Systems.* Heidelberg, Germany: Springer-Verlag.

VAISSIÈRE, J. (1989). On automatic extraction of prosodic information for automatic speech recognition system. *Pages 202–205 of: Proc. 1st. European Conference on Speech Communication and Technology*, vol. 1. Paris.

VAN BERGEM, D.R. (1988). The first step to a better understanding of vowel reduction. *Proc. Institute of Phonetic Sciences*, **12**, 61–75.

VAN BERGEM, D.R. (1994). A model of coarticulatory effects on the schwa. *Speech Communication*, **14**(1), 143–162.

VAN SANTEN, J.P.H. (1993). Compensation for vowel coarticulation: A progress report. *Pages 35–40 of: Proc. ARPA Workshop on Human Language Technology, Plainsboro, New Jersey, 21–24 March 1993.*

VAN SUMMERS, W. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analysis. *Journal of the Acoustical Society of America*, **82**(3), 847–863.

VERBRUGGE, R.R., & SHANKWEILER, D.P. (1977). Prosodic information for vowel identity. *Journal of the Acoustical Society of America*, **61**(Supplement 1), S39(A).

WAIBEL, A. (1988). *Prosody and Speech Recognition.* London: Pitman.

WESTIN, K., BUDDENHAGEN, R.G., & OBRECHT, D.H. (1966). An experimental analysis of the relative importance of pitch, quality, and intensity as cues to phonemic distinctions in southern Swedish. *Language and Speech*, **9**, 114–126.

WIGHTMAN, C.W., & OSTENDORF, M. (1991). Automatic recognition of prosodic phrases. *Pages 321–324 of: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. Toronto, Canada.

WIGHTMAN, C.W., & OSTENDORF, M. (1992). Automatic recognition of intonation features. *Pages 221–224 of: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. San Fransisco.

WIGHTMAN, C.W., SHATTUCK-HUFNAGEL, S., OSTENDORF, M., & PRICE, P.J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, **91**(3), 1707–1717.

ZWICKER, E., & TERHARDT, E. (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.