AUTOMATIC PROSODIC ANALYSIS FOR COMPUTER AIDED PRONUNCIATION TEACHING

Paul Christopher Bagshaw

A thesis submitted for the degree of
Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

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 Aussprachеunterricht

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 (*FØ*) 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 *FØ* ist notwendig als anfänglicher Prozeß in der automatischen prosodischen Sprachanalyse. Die Probleme werden addressiert, die auftreten, wenn man versucht *FØ* so festzustellen, daß minimale Fehler in der prosodischen Analyse auftreten, da die *FØ*-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 (*FØ*) 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 *FØ* initialement pour l'analyse prosodique automatique de la parole. On considère les problèmes posés par la nécessité de déterminer *FØ* 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

Declaration of Originality

The material contained within this thesis and the work reported within are solely the original work of the author. When information or ideas have been taken from other texts or derived from personal communication with other researchers, the source of information has been clearly stipulated. This thesis has not been submitted for a degree or professional qualification at any other academic institute.

Paul C. Bagshaw

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

List of Figures

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

List of Tables

Glossary of Technical Abbreviations

♀	female speaker.
♂	male speaker.
"$-$, $\backslash$, $/$, $\vee$, $\wedge$"	$F\emptyset$ trajectory descriptors.
Δf	$F\emptyset$ 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 $F\emptyset$ in schema.
"ap"	accented syllable according to pitch accent decision filter.
$C_{1,2,3}$	features of formant frequency trajectories.
CFD	cepstral-based $F\emptyset$ 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.
eSRFD	enhanced super resolution $F\emptyset$ 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.
$F\emptyset$	fundamental frequency.
$F1$, $F2$, $F3$, $F4$	formant frequencies.
FBFT	feature-based $F\emptyset$ tracker.
FDA	$F\emptyset$ 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"$\cup$"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 ($F\O$) 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 *FØ* extraction from propagating into the prosodic analysis. An enhanced super resolution *FØ* determination algorithm (*e*SRFD) is proposed in Section 7.1.

Section 7.2 presents a comparative evaluation of the *e*SRFD algorithm and the *FØ* determination algorithms summarised in Section 3.2. The evaluation is performed by comparing *FØ* 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 *FØ* determination algorithm are shown to radically reduce the number of gross *FØ* 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 *FØ* contour generated by an *FØ* determination algorithm in a way which further reduces the occurrence of errors. 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 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 *e*s) 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.

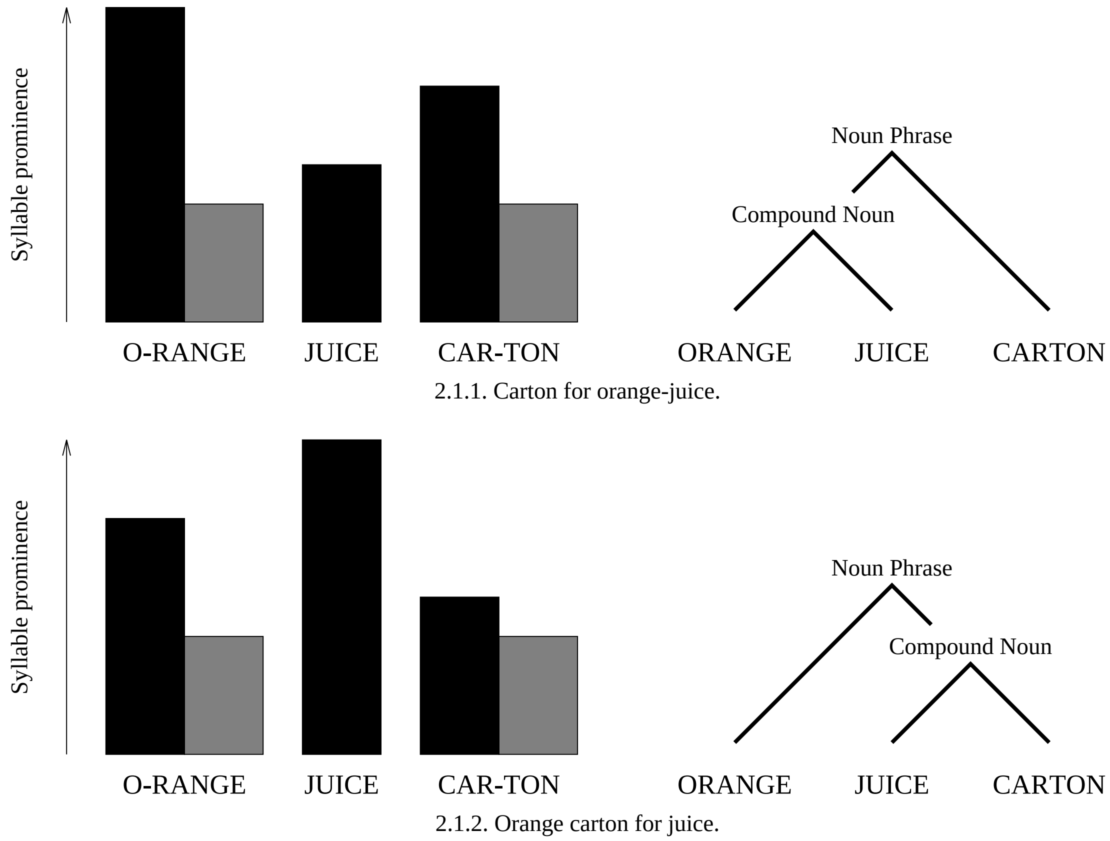

Figure 2.1: Syntactic and phonetic variations of *orange juice carton*.

word *juice*) meaning, ‘a carton for orange-juice’; and “**o***range* **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* /ˈa ŋ k o r a/ 'anchor', /a ŋ ˈk o r a/ 'again'; *compito* /ˈk o m p i t o/ 'task', /k o m ˈp i t o/ 'poised'; *turbine* /ˈt u r b i n e/ 'whirlwind', /t u r ˈb 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 ˈb 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* /ˈt ɜ b aɪ n z/, with the lexical stress on the initial syllable, and tends towards the pronunciation of the Italian word /ˈt 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.

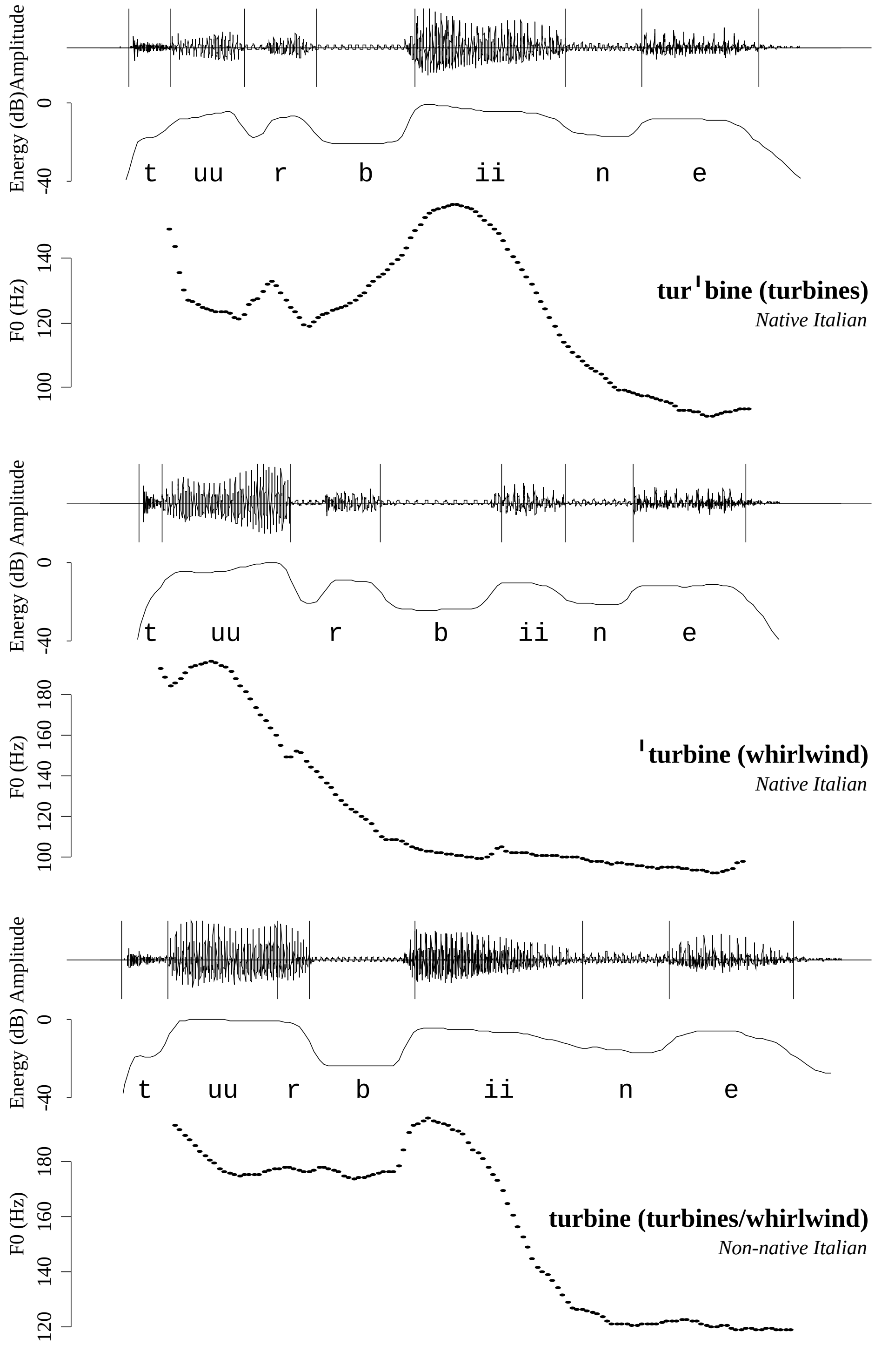

Figure 2.2: Phonetic realisations of the Italian word *turbine**.

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 ˈb 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 ˈb i n e/ 'turbines' and /ˈt 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 ˈ). 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 ˳) 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.

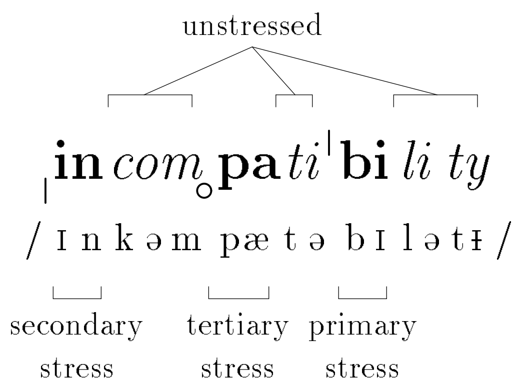

Figure 2.3: Degrees of stress in *incompatibility*.

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.

$$\textbf{Tone unit} = (\textit{Prehead})\ (\textit{Head})\ \textbf{Nucleus}\ (\textit{Tail})$$

For example,

Accented syllables

She was ˌgoing ˌhome to ˈsleep this afternoon

Prehead *Head* *Nucleus* *Tail*

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 intonation 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.

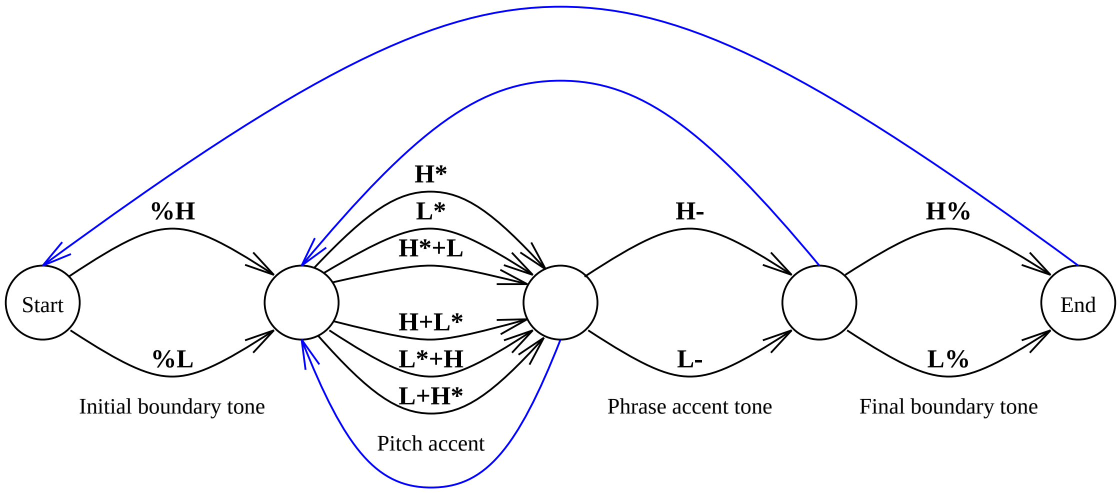

Figure 2.4: Finite-state grammar for **H/L** tone sequences (Pierrehumbert, 1980)

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."

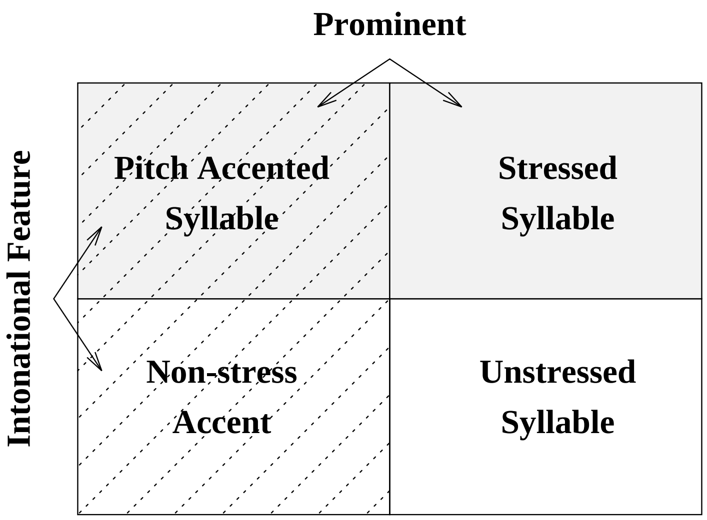

Figure 2.5: Prominent syllables and intonation features

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.